Validation
of
Risk Assessment
in
Auditing

ISBN 90 8659 055 1

VRIJE UNIVERSITEIT

Validation of Risk Assessment in Auditing

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Economische Wetenschappen en Bedrijfskunde
op maandag 18 december 2006 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Gerrit Benjamin Broeze

geboren te Edam

promotor: prof.dr. P.G.W. Jansen
copromotor: prof. J.H. Blokdijk RA

CONTENTS

Preface

An auditor assesses the risk that financial accounts, offered to him (her) for certification of the truth and fairness, contain a material error. He (she) may rely on this assessment by reducing his (her) substantive audit effort when the risk is deemed to be low. So it is reasonable to expect that this risk assessment is consistent with the outcome of his audit in terms of the size of the error found and in terms of estimates of the real risk that can be based on this error. By engaging in the research reported in this thesis (in 1996) I tried to fill a gap in the knowledge on this issue. Many persons and organisations have contributed to its realisation. I am very grateful to all of them.

In the first place I mention the 'Algemene Rekenkamer', the Netherlands Court of Audit (NCA), I had the pleasure to work for from 1986 until 2005. These nineteen years I spent as a statistical consultant, in a context where there is place for questions, curiosity, amazement. This not only because of the scepticism which, maybe, is a necessary condition for a proper execution of the task of a court of audit, but also because of interest in improving the audit profession. So I was given the opportunity to engage in my research, not only by being given the time, but also by numerous discussions with my colleagues, almost without exception highly interested.

In the second place I mention the Limperg Instituut (LI), which gave me an appointment for a day a week from 1996 through 2004. In this time I did the research, reported it to the organisations which participated by giving access to their data and started to write this thesis.

In the third place I mention the organisations which were willing to give me the necessary data. Without their disinterested cooperation, this research would have been impossible. Our agreement on anonymity forbids mentioning names, also of people in these organisations who have greatly contributed to the ideas in this thesis.

I mention the Steering Group Statistical Audit, in which the many discussions on the use of sampling and the validity of risk assessment were a great stimulus to start and continue this research.

I mention the "Kenniskring Steekproeven" (the 'Statistical Auditors of the Round Table'), a peer group of auditors from governmental audit organisations, interested in the application of statistical sampling in auditing. They bore my imperfect knowledge of accounting and auditing, my still confronting them with my questions and half bred ideas regarding (improvement of) risk assessment and stimulated me to continue my research, by always showing their interest and stressing the relevance of my questions.

I mention with pleasure and gratitude the many discussions with Henk Kuenen, Berrie Zielman, Martin Dees, Jurrie Vos, Marion van Dam, Nur Acardag and many other colleagues of the NCA, with Fred Drieënhuizen, Chris Hibbitt and Marcelle Weisfelt of the LI, with Paul van Batenburg of Deloitte and Touche, who spent many hours discussing with me the (im)possibilities of validating risks and other topics in statistical auditing, with Martien van Zuijlen of the Radboud University, who has such a gentle way in getting people to clarify their problem, with Hein Kloosterman, Niek de Jager, Ruud Veenstra, Lucas Hoogduin, Jacques de Swart, members of the steering group and also partners in some subprojects, with Roelof Helmers (CWI), who also spent much time with me in exploring possible relations in my risk indicators data set and with Hans Moors, who gave feed back on the statistical parts of the manuscript.

I mention Bianca Snel and her supervisor Karma Dajani of the Rijksuniversiteit Utrecht, who gave a major contribution to the simulation studies used in chapter 4. Without their effort a valuable and necessary part of my research would have been impossible.

I mention the contributions of Berrie Zielman, Lauw Simonse, Pauline de Groot, Frans Vis and Nur Acardag, all employed by the NCA, to the work in the chapters 8 and 9.

I mention Robert Mureau of the KNMI (the Royal Netherlands Meteorological Institute), who introduced me into the 'Ensemble Prediction System'.

I mention Luc Quadackers, who greatly helped me collecting literature.

I thank Jack Franklin for being so kind to read the whole manuscript and correct the English where necessary. (It was never meant to conceal the fact that I am not a native speaker).

I thank the members of the reading committee for spending their time in reading my manuscript and providing me with valuable comments; these certainly led to improvement of my thesis.

Last but not least I thank my (co)promoters, Paul Jansen and Hans Blokdijk, for always seeing the good intentions in drafts for a new chapter, spaghetti westerns I offered them to eat and digest. Still they always had valuable suggestions and their patience never failed.

This work was always intended to be 'just work', to be done during the normal daily working hours. But of course it lasted longer than planned, so that after my partial retirement I still had to spend much of my 'own time' to finish the thesis in a reasonable time. So in the end the work intruded into my private life, hindered plans we had and thus made an appeal to my wife's patience. I thank her for having that and giving me the support to finish this work. And most important of all: I thank her and my two daughters for their love.

Edam, 30 September 2006

Chapter 1: Introduction

In the audit risk approach, the assessment of the occurrence risk is crucial. This thesis intends to validate this assessment with the observed error rate and related criteria. The motivation for this goal is given (section 1.1) followed by an introduction into the audit risk model, which is the context for risk assessment (section 1.2). This introduction includes a closer look at the concept of audit risk (1.2.6).

1.1 Quest for the real risk

Common audit practice uses risk analysis (*assessment*[1] of the risk of not discovering an error of a material size, the "*audit risk*"[2]) in two ways: (1) it guides the auditor in the choice of his/her audit objects (in this thesis: "*the primary allocation*"), (2) when the risk is assessed to be low, risk analysis may replace a part of the *substantive tests of details* or other substantive procedures (in this thesis: "*the secondary allocation*") and sometimes even all of these tests. This practice is part of the so-called "*audit risk approach"*. I first met it as a statistical consultant at the Netherlands Court of Audit (starting 1986). It is still observed to be common practice in a statement of the Panel on Audit Effectiveness in POB (2000, p.33): *"The major reason for the auditor's risk assessment activities is to provide a basis for determining the nature, timing and extent of substantive tests to be performed to provide a reasonable assurance the auditor needs about the reliability of the assertions that are embedded in the financial statements."*

As a formalisation of this approach, the audit departments of the Dutch ministries base the extent of substantive testing on a table which gives the reliability still to be generated by an audit sample, with IR, ICR and RAR as arguments (see HCDAD, 1997). All departments (may) use this table, regardless the specific qualities of their risk assessment. The uniformity in the use of the tables can only be justified when the audit approaches and the risk assessments of all users of these tables are also uniform. Partly this uniformity is realised, because all departments use the same handbook, maintain professional exchange and share education. But in many cases differences are visible without much effort. And, also with risks assessed in accordance with professional standards, there is no guarantee that an assessment "low" by one auditor represents the same risk as the same assessment by another one. The same or even more severe lack of guarantee applies to assessments compared between departments.
This lack of uniformity is confirmed in the report of the Panel of Audit Effectiveness (POB 2000, p 34, 37) which reports large variability in sizes of samples designed by different members, even of the same audit department of an audit firm, who come to the same formal risk assessment.

These observations led to questions like:

- Is the audit risk approach (formalized as the *audit risk model,*(ARM, see among many others: POB, 2000 (pp 175-179), NIVRA, 1989, Arens & Loebbecke, 1997 (pp 257-264), ISA 200 par. 16) justified, insofar it allows the auditor to replace

[1] "Assessment" is a term which is part of natural language; it is that "primitive" that the IAASB Handbook 2004 does not think it necessary to define it. Because of the central role it plays in this thesis, we still give a description of it in the glossary.
[2] Italicized words refer to concepts that are defined in the glossary.

substantive tests of details, or other substantive procedures, by the information from risk analysis?
- Is this practice validated in some way or another?
- Has the above mentioned table of the governmental audit departments in the Netherlands been validated?

An answer to these questions should come either from earlier empirical research, or from research by the audit organisation itself. Our chapter 3 will show that research fails to give a satisfactory answer. Moreover the audit organisations I discussed these questions with did not try to do the necessary empirical research, so I decided to start my own empirical research, from a question that can be seen as generic for these validation/ justification questions:

Generic question
"Does risk assessment by an auditor represent the 'real risk' of a material error?".

But because of difficulties concerning a concept like "real risk" I decided to investigate something more simple: "Is there a satisfactory relation between risk assessment on the one hand and error rate, and some measures derived from it, on the other?". This is the basic research question and it will be leading for the investigations and further be elaborated in this thesis. The basic question is formulated as follows:

Basic research question:
"Is there a satisfactory relation of the assessed risk with the error rate as found in an audit, or with other measures, derived from it, that may be seen as an indicator for the risk of a material error?".

I wanted to deal with this question with real life data as a source, so I decided to perform a field study, with 'real risk assessments' and with 'real error rates'. Chapter 5 will explain this in more detail.

1.2 The audit risk model in auditing

The aim of auditing is to give an opinion on the quality of a financial statement (see among many others: ISA 200, par. 2). In principle, this opinion is of the form: "My opinion is that the annual accounts do not contain an error of material size, except for an acceptable risk that my opinion may be false because of not having detected a material misstatement ".

In this opinion "*material misstatement*" or "*material error*" means a misstatement of a size as to influence the economic decisions of users of the annual accounts (see e.g. Handbook IAASB 2004, p.141). In practice the choice of a level of materiality is governed by conventions, which are the result of experience, negotiations of users and auditors and public opinion. In Dutch governmental administrations, a level of 1% of the account size is usually considered 'material', with private companies 5% of the profit is not unusual. The risk of giving a false opinion is called the *audit risk* (AR, see ISA 400 par 3); in general a maximum for the audit risk of 5% is deemed to be acceptable.

In order to deal with the audit risk, the so-called "audit risk model" is in use. The crux of this model is twofold:
1. it decomposes the audit risk into two components: the occurrence risk and the detection risk;

2. it is hoped and assumed that the assessment of the occurrence risk provides so much assurance as to the absence of a material error, that the total cost of the audit (the detection activities included) will be less than the costs of an audit purely based on detection activities (see again POB 2000, p.33).

.

The two components of the decomposition of the audit risk can be introduced as follows:

For the first component the leading question is: “What is (prior to the audit), the risk of the existence of a material misstatement or error in the accounts?” The auditor will try to assess this *occurrence risk*, OR.

For the second component the leading question is: “What is the risk of not detecting a material misstatement, although it is in the accounts?” The auditor will try to control this *detection risk,* DR, by a proper planning of the audit.

Both OR and DR are decomposed in their turn: OR in the *inherent risk* and the *control risk* (see 1.2.1)*,* DR in the *risk of analytical review* and the *sampling risk* (see 1.2.2)*.* Clearly the audit risk will be a function of the occurrence risk and the detection risk. In this thesis the assessment of the occurrence risk will be the object of validation. We will go deeper into its assessment.

1.2.1 The occurrence risk

In assessing the occurrence risk (OR) the auditor will decompose the factors to be assessed into two types: (1) those regarding the *inherent risk* and (2) those regarding the *internal control risk* or *control risk*. He will see OR as a function of the inherent risk (IR) and the control risk (CR):

$$OR = f(IR,CR).$$

ISA 400 (par. 4) gives the next definition for inherent risk:

Definition of inherent risk:
“Inherent risk” is the susceptibility of an account balance or class of transactions to misstatement that could be material, individually or when aggregated with misstatements in other balances or classes, assuming that there are no related internal controls.

Essentially, the inherent risk regards the risk of a material error due to factors that cannot be, or are not controlled by the business and administrative processes. It can be seen as the risk originating from the context, like environment, type of activity of the business, etc.. ISA 400 (par. 12) states factors on which the inherent risk is dependent, some of them being:

- the integrity of management
- managements experience and knowledge
- changes in management
- pressure on employees
- the nature of the business

Many of the business and administrative processes are designed to control the risks given with the context. The control risk regards the risk due to the fact that these business and administrative processes cannot control every possible mistake or misstatement. The leading question in assessing this risk is: will the business processes, in particular the administrative processes and related controls, prevent material errors? And if not, will these controls detect and correct them? The quality of these processes determines the internal control risk, ICR, or (synonymously) control risk, CR, defined as ISA 400 (par. 5) does:

Definition of control risk:
"Control risk" is the risk that a misstatement, that could occur in an account balance or class of transactions and that could be material individually or when aggregated with misstatements in other balances or classes, will not be prevented or detected and corrected on a timely basis by the accounting and internal control systems.

ISA 400 summarises components of the 'control environment' (par 8) and goes into the question how the auditor can get an "understanding of the business" and how control risk can be assessed by means of tests of control. We mention:

- verification that a transaction has been authorised
- inquiries about and observation of internal controls which leave no audit trail
- reperformance of internal controls, for example reconciliation of bank accounts (par 30)

The auditor investigates the processes and related controls as to their design and their existence as well as to their operation; their operation is assessed by means of *tests of control*. From these assessments and analyses the auditor assesses the control risk (CR). Together with the inherent risk (IR) it forms the occurrence risk (OR). So OR can be seen as a function of inherent and control risk.

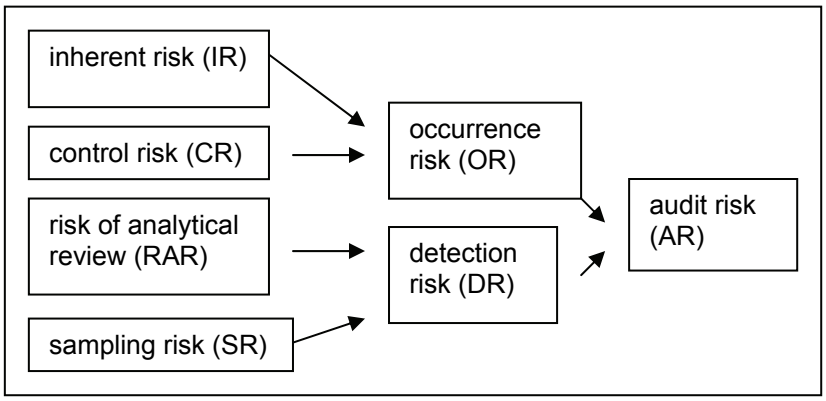
This function is sometimes viewed as just a function, without further specifications, and otherwise as the multiplication of IR and CR, (see also 1.2.5). Many auditors do not separately consider IR and CR, because in their assessment some factors can both be accounted for in IR and in CR (see e.g. Touw & Hoogduin (2002), pp. 13,14).

This thesis focuses on the occurrence risk, because of the practice just mentioned, and because the combined inherent risk and control risk are the risks an auditor has to deal with in his planning and execution of the audit. This occurrence risk is taken as a fact for the auditor, once he starts his detection activities, by means of *substantive procedures.*

Definition of substantive proecedures
"Substantive procedures" are concerned with amounts, aimed at obtaining audit evidence to detect material misstatements in the financial statements, (see ISA 530, par.17).

The decomposition of the audit risk can be depicted as in figure 1.1. The decomposition of the detection risk will be discussed in 1.2.2.

Figure 1.1: Decomposition of the audit risk



Our discussion of the audit risk enables us to formulate the generic question in section 1.1 more precisely as:

Does the occurrence risk, as assessed by the auditor represent the 'real risk' of a material error?[3]

1.2.2 The detection risk

Figure 1.1 shows the decomposition of the audit risk in occurrence risk and detection risk. In this subsection we discuss the detection risk.

Definition of detection risk:
"Detection risk" is the risk that substantive procedures will not detect a misstatement that exists in an account balance or class of transactions and that could be material, individually or when aggregated with misstatements in other balances or classes. (see ISA 400 par 6).

The detection risk is broken down into the "risk of analytical review" (RAR) and the "sampling risk" (SR):

$$DR=f(RAR,SR).$$

Definition of risk of analytical review:
The "risk of analytical review" (RAR) is the risk that analytical procedures will not result in the detection of a misstatement that exists in an account balance or class of transactions that could be material, individually or when aggregated with misstatements in other balances or classes (see ISA 520 par. 10-15, where only *reliance* on analytical procedures is discussed). [4]

Analytical procedures play a role both in the planning phase and in the completion phase of the audit, as we will see in the next subsection. If possible the auditor also decides on the risk of not detecting a material error by way of a (statistical) sample, more comprehensively defined as:

Definition of sampling risk
"Sampling risk" (SR) arises from the possibility that the auditor's conclusion, based on a sample, may be different from the conclusion reached if the entire population were subjected to the same audit procedure." (see ISA 530 par. 7).

Sampling risk plays a role in the planning and in the completion stage, as we will also see in the next subsection.

---

[3] For an error found, the auditor will investigate its kind, cause and consequences. This may also lead to the conclusion that the error regards fraud. So our research will also apply to the risk regarding fraud. Sometimes an investigation is aimed specifically at discovering fraud and then extra tools are used.

Our research regards the regular audits, which allows conclusions as to predictability of fraud, but does not apply to these extra tools.

[4] It is a remarkable that neither ISA nor SAS explicitly define the risk of analytical review, where at the same time it has an explicit place in HCDAD (1997) and in handbooks of various big firms. And it has a place in the practice of auditing.

1.2.3 Planning of the audit

Having assessed the occurrence risk, as discussed, by observing and making an analysis of the business, its context and its (administrative) processes, the auditor can start to plan the audit. In this planning stage, his knowledge of the inherent risks and the quality of the administrative procedures and internal controls will play a key role in allocating the audit resources, the primary allocation, as we introduced it in section 1.1 (figure 1.2, arrow 1a).

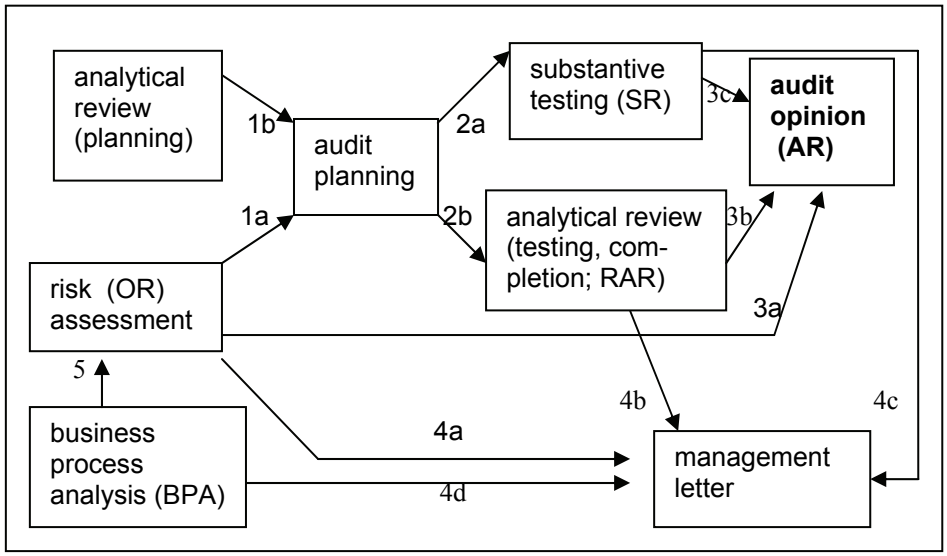
In accordance with ISA 520 (par. 2, 7) the auditor will apply analytical procedures, to get a better view where other audit procedures should be applied. So next to the assessment of OR, analytical procedures play a role in the primary allocation of audit resources (figure 1.2, arrow 1b). They also play a role in the testing and completion stage of the audit and at other stages where analytical review is more effective or efficient than tests of details (arrow 2b). Both the analytical activities in the planning and in the testing and completion stage imply the "*risk of analytical review*"[5].

Once the primary allocation has been decided upon - the *audit objects*, the parts of the annual accounts and their context that will be subjected to the ongoing audit have been determined - a set of detection activities is planned for these parts. These detection activities will consist of substantive procedures: substantive testing (also tests of detail, arrow 2a) and/ or analytical review (arrow 2b). A minimum of substantive procedures is always required (see ISA 330, par 49). Dependent on the nature of and the extent to which these procedures are performed, the auditor runs a detection risk (DR), the combination of sampling risk (SR) and risk of analytical review (RAR). As a rule, he plans the substantive procedures to an extent as resulting in an audit risk of at most the maximum that he (and his auditee and society) is prepared to accept (mostly 5%). This makes the planned maximum for the detection risk dependent on OR. To complete the picture, the audit activities lead to an audit opinion, with a planned maximum for the audit risk (AR). In summary: this audit risk is controlled by the assessed occurrence risk, the planned risk of analytical review and the planned sampling risk (figure 1.2, arrows 3a, 3b, 3c).

As a rule there are many observations and findings that are worthwhile to be reported to management, also when these do not directly influence the audit opinion. These findings may result from every audit activity and lead to the so-called "management letter" (arrows 4a, 4b, 4c), in which these findings are reported and management is given some advice where necessary or useful in the opinion of the auditor. As far as business process analysis (BPA) is performed, this will certainly contribute to this management letter (arrow 4d). BPA may also affect the assessment of OR (arrow 5). BPA will be discussed in subsection 1.2.7. Figure 1.2 depicts the relations we just discussed.

[5] With this decomposition of DR, there is an inconsistency in the audit risk model: RAR is seen as component of DR, but to the extent that analytic review plays a role in the audit planning, the associated risk is not defined. In this thesis RAR will not play a role, because it is not a part of OR.

Figure 1.2: Audit planning in accordance with the audit risk model.



In addition to the points discussed above we should make the following remarks.

Remark 1: The audit opinion rests on three types of activities.
In the discussion above this was already mentioned. But it is interesting enough to stress it in an explicit remark. The activities, also depicted in figure 1.2, are: (1) substantive testing or other substantive procedures, (2) analytical review and (3) the outcome of risk analysis itself: the assessment of the occurrence risk.

Remark 2: Substantive procedures give the most direct evidence.
Substantive testing and some forms of analytical review, give the most direct evidence as to the truth and fairness of the account under audit. Individual *book-values* are tested for their correctness. The aggregate of errors found in the tests gives an estimate of the size of the total error in the audited account (arrow 3c). In principle, the level of risk associated with the audit opinion is largely determined by the size of the sample that is taken from the account. In many cases this sample is selected on a random basis, allowing sound statistical inference on the error rate. This inference as a rule, is done by giving a *point estimate* (*most likely error, MLE*), together with a *confidence upper limit* (*upper error limit UEL*) at a *confidence level* that corresponds to the desired "level of assurance". "*Assurance*", also *"audit assurance"* is the complement of audit risk: when, for instance, the audit risk equals 10%, the audit assurance equals 90%. When substantive tests are the only basis for the audit opinion, statistical laws exactly produce the level of audit assurance, or equivalently audit risk, associated with the audit opinion, provided the tests were selected on a statistical basis. The secondary allocation especially determines the size of the audit samples.

Remark 3: Sampling risk is used in the planning and in the evaluation.
In the planning stage the sample is planned at a size such that a maximum for this SR ('the 5%') will not be exceeded, when the actual error is equal to the materiality. Once the sample has been audited, the likelihood can be calculated that the real error in the population (= the account under audit) may exceed the materiality, given the error found in the sample. This likelihood is also seen as the sampling risk. This post hoc SR may be both larger and smaller than the planned (allowed) maximum. When it does not exceed the this maximum, the sample results allow an unqualified opinion. In this thesis this post hoc SR plays an important role as a validation criterion (see Ch. 2)

Remark 4: The occurrence risk is the responsibility of the auditee, the detection risk is the responsibility of the auditor.

This distinction offers another way to think of the occurrence risk and the detection risk. It should be kept in mind, however, that the assessment of the occurrence risk is the responsibility of the auditor.

Remark 5: For this thesis assessed risk and error are the relevant outcomes.
For this thesis the relevant outcomes of the audit activities are the assessed OR and the assessed error in the annual accounts on which assessment the audit opinion will be based. Not all activities of figure 1.2 are aimed at these two outcomes; they serve to give context to our analyses of the error rate in its relation to risk assessment.

1.2.4 Risk analysis replacing substantive testing

In the logic of the ARM, risk analysis is supposed to yield information on the quality of the annual accounts. It leads to a practice where exact formal probabilities are calculated, partly based on the judgmental assessments of risk analysis. And as a consequence, the model assumes that the extent of the substantive testing (synonymous with 'substantive tests of details') may be decreased if the occurrence risk is assessed at "low" or "medium". In the start of this section (1.2) we mentioned that this assumption can be seen as one of the main reasons for risk analysis (see for instance POB 2000, p.33, JWG 2000, p36, 42, NIVRA 1989).

This combination of judgments and exact statistics is very common in *Bayesian statistics.* There the prior distribution plays a role parallel to that of the occurrence risk in auditing: both methodologies provide a way to quantify the degree of belief the auditor has that a material error will be absent. In the field of statistics, there are doubts on the validity of such prior information especially when modelled by a prior distribution. These discussions regard two aspects: is there really information that applies to the probability that is to be calculated, and is it possible to transform the information into the probability in such a way that the calculus gives a valid result. (see e.g.: Lee 1997, pp. ix,x,xi, Box & Tiao 1992, p 12, Novick & Jackson 1974, pp 145,146, Sennetti 1995, Johnstone 1995, Loebbecke 1995). To circumvent these doubts, in Bayesian statistics a so-called 'non-informative' or 'reference prior' is often chosen. Such a prior distribution expresses the beliefs of someone who, a priori, had no strong beliefs about the parameter in study, and therefore is such that the posterior distribution is almost fully dependent on the likelihood of the sample result (See Lee 1997, p.45)

The position of auditing in these discussions is relatively favourable: risk analysis provides much prior information that applies to the situation under audit and that also relates to some probability; obviously the ARM is built on this information. So if there are fields where Bayesian statistics is applicable, auditing surely belongs to these fields. Nevertheless, in the audit context too the transformation of information into a probability is difficult and questionable. In that respect it has similar difficulties as application of the ARM. But where there is such an abundance of prior information, it is worthwhile to investigate the validity of this transformation, and not to take refuge into non-informative priors. This latter approach would boil down to not taking prior information into consideration at all, for the level of assurance associated with the audit opinion.

In this thesis we will not elaborate on the Bayesian approach as such. We will only make use of the approach when we calculate the sampling risk (see remark 3 above and chapter 4, where we will complete this introduction on Bayesian statistics to a minimum, needed for this thesis; see also the glossary). And our most important suggestion for improvement of the use of risk assessment includes the use of Bayesian statistics (see chapter 10).

1.2.5 The audit risk model

Our discussion of the audit process and the place of risk analysis leads to the complete audit risk model (ARM). It appears in two forms.

1. The first form is that of a multiplication : AR = IR x CR x DR (see a.o. SAS 47, Arens and Loebbecke 1997, p. 257, NIVRA 1989),
   or with the detection risk broken down: AR = IR x CR x RAR x SR (see a.o. HCDAD 1997).
   The ARM in this form is assumed to be multiplicative. This assumption has caused extensive discussions because the independence of the various risks, which is a necessary condition for this model to be valid, can be questioned (see for instance Broeze et al, 1991, Schilder 1991).
2. This discussion has led to a second form for the ARM: AR = f (IR, CR, RAR, SR). This form is chosen merely to imply that the auditor is aware of the fact that a precise functional form for the audit risk and its alleged determinants is difficult to find (or easily criticized) but that this does not alter the fact that there is a relation between these risks.

The second form also formalizes a frequently used argument in the discussion on the merits of the multiplicative form: the ARM is just a heuristic model, without claims on precise mathematical properties.

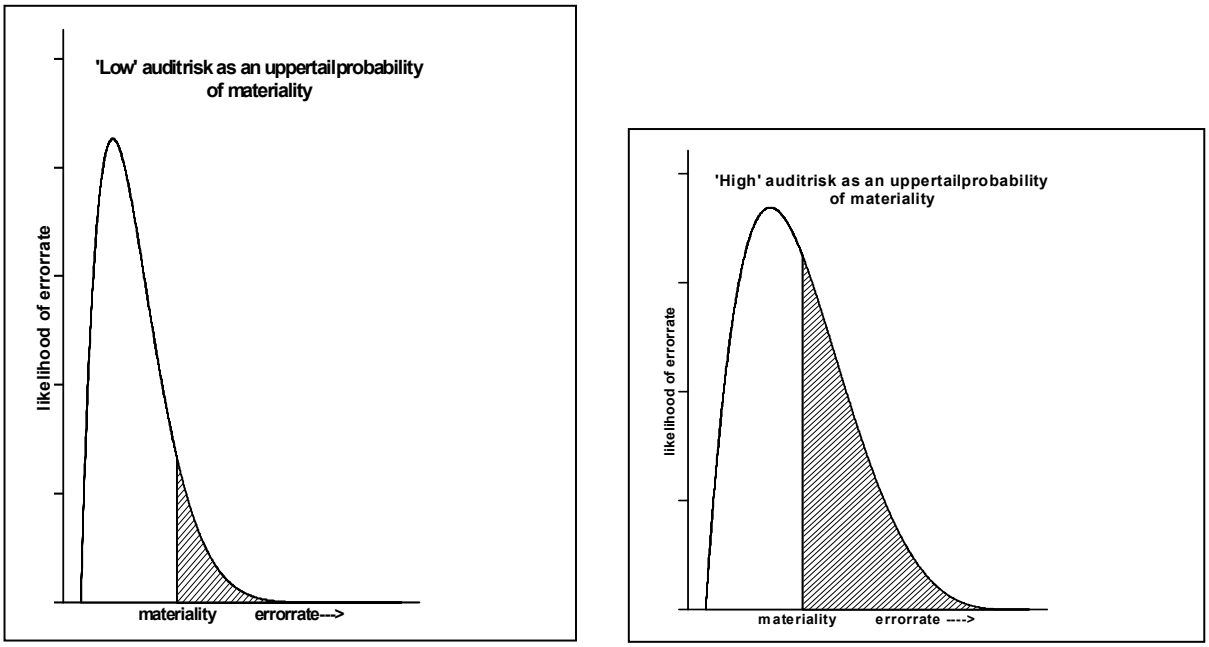
The table mentioned in section 1.1, used by Dutch governmental audit departments, is based on the second form. It gives a level of confidence still to be generated by the audit sample, with IR, CR and RAR as arguments (see the handbook HCDAD 1997 325.38). The assigned confidence level, in combination with the assessed IR, ICR and RAR is supposed to establish an audit risk of at most 5%. The table is not based on a multiplicative model, because the combination of IR='low' with ICR='high' assigns a confidence level different from the combination ICR='low' with IR='high'. For many more combinations the assignment is not commutative, as it would be in a multiplicative model.

1.2.6 Audit risk as a property of the distribution of possible error rates

So far we have treated audit risk as a function of other risks. Now we wonder in an exploratory way how the meaning of audit risk could be formalised as a property of the distribution of possible errors, before we will analyse possibilities for validating the occurrence risk in chapter 2.

We realize that risk and error rate are related, but that "risk" is more than error rate: it also includes materiality and probability. We see this in figure 1.3. This figure depicts the possible error rates in an annual account along the abscissa, the "error rate-axis". The curve depicts the likelihood of all these possible error rates: the higher the curve above an error rate, the more likely this error rate and the error rates close to it. Obviously, this likelihood can be a product both of risk analysis and of sample results. The area under the curve between two possible error rates depicts the likelihood of these error rates. For instance the area under the curve to the right of the vertical line depicts the likelihood of the possible error rates to the right of this line. If the position of this line represents the level of materiality, the shaded area represents the likelihood of a material error. So the shaded area depicts the audit risk, in case an auditor would give an unqualified opinion on the annual accounts to which this picture applies, without doing further auditing. In a more formal sense, audit risk can be seen as the likelihood of errors of a material size (when an unqualified opinion is given). We will extend this view in chapter 2.

Figure 1.3: materiality and audit risk



1.2.7 Business process analysis

By the end of the nineties of the last century an evolution in the application of the audit risk approach resulted in an approach in which the whole of business processes and the context of the business were analysed as to the possibilities and risks of (not) obtaining the objectives of the business. This way of dealing with a business, its continuity and prospects, its accounts and their audit, is known as "*business risk analysis*", "*business risk approach*" (see JWG 2000 pp. 3,5), "*business risk audit model*" (Ellifsen et. al. 2001) or "*business process analysis*" (see Van Leeuwen & Wallage 2002). Business process analysis (BPA) is meant to get an image of a business' strategic position, the quality of the business' products and processes and of the administrative processes in particular. This all is meant to assess the strategic situation, and get an image of the business itself.

BPA is also meant to be relevant for the audit of the business' annual accounts: it is even expected that BPA gives a better view of the occurrence risk than the ARM, because the auditor not only has an image of the accounting process and its related controls, but also of all other business processes and in the way they interact with the accounting process. (see e.g. Van Leeuwen & Wallage 2002, JWG 2000 p 33).

In 1998 several international accounting and auditing organisations installed a "Joint Working Group" which was to investigate recent developments in audit methodology, in particular the extent to which BPA was in use with audit firms and governmental audit organisations. This Joint Working Group (JWG 2000, p.10) has the same expectation as to the effectiveness and efficiency of BPA as have Van Leeuwen and Wallage, especially for issues concerning accounting estimates, going concern and management fraud. The Joint Working Group mentions nine reasons for adopting the business risk approach (JWG 2000 pp. 35-37), among which audit effectiveness, audit efficiency, control of engagement risk (the risks for an auditor due to engagement in an audit) and client service.

In this thesis we focus on the audit risk model, because the data we have all were produced in the context of the audit risk model. Therefore henceforth business process

analysis will not be paid attention to. Only in chapter 10 we will discuss the question to which extent our findings are relevant for those who use business process analysis.

1.3 Aim of the study

The aim of this study is, to make some steps in the quest for the 'real risk', mentioned in section 1.1, finding evidence whether substantive testing may be replaced by risk assessment as it is done in practice, and if so, under which conditions this will work. As risk assessment plays such as central role in auditing, the answer may have practical implications.

A necessary condition for this practice of replacement to be justified is a positive answer to the basic research question in this study:
"Is there a satisfactory relation of the assessed occurrence risk with the error rate as found in an audit, or with other measures, derived from it, that may be seen as an indicator for the risk of a material error?"
So we can formulate the aim as follows:

The aim of the study is to get insight into the relationship between risk assessment and error rate, in order to get an answer to the question whether risk assessment can be used as a replacement for substantive testing.

1.4 Participating organizations

In chapter 5 we will go deeper into the design of the study, but an essential feature of it should be stressed now: without the willing participation of many organizations, this study would have been impossible. We agreed to report our findings in a way that they cannot be associated with one of the participants in the study. This is the reason that we only mention them in this chapter in a form of disguise and only will refer to them in the chapters 6 through 9 again in the same way as to prevent recognition.

The participating audit organisations were:
1. A national court of audit from one of the countries of the European Union
2. The audit departments of eight Ministries of the Dutch government
3. Two private audit firms (of the 'big five') in the Netherlands,

Data were collected regarding audits on annual accounts from 1995 through 2001.

1.5 Structure of the thesis

In order to answer the basic question, we have structured this thesis as follows.
1. This introduction gives the motivation and basic question of this thesis. Also a short introduction on the audit risk model and basic definitions are given.
2. The second chapter will treat the basic question as the problem of validity of risk assessment with regard to four validation criteria:
    1. the error rate
    2. the 'audit position'
    3. the sampling risk
    4. the conditional distribution of the error rate.

Each criterion will be discussed as to their quality as a validation criterion.

3. The third chapter will give a review of relevant literature. It will speculate on the possibilities of humans to assess risks, in which speculations the work of Tversky and Kahneman will serve as a guide. Next to that it will report findings related to our basic question; the picture that results is neither a cause for optimism nor for pessimism.
4. The fourth chapter will justify the use of the beta distribution as a probability law for the sampling risk. We need this justification, because in our study we had to calculate sampling risks on the basis of limited information. It appears that the beta distribution is a model in which this information is sufficient for the required calculations.
5. The fifth chapter will deal with the design of this study. It will justify our choice for a field study.
6. The sixth chapter will give validation results for the audit risk model for eight audit organisations, both private and governmental. Private and governmental organisations show no clear differences. At the same time, the results vary strongly over organisations and therefore ask for more evidence that risk assessment has the required validity. This evidence is sought in the chapters 7, 8 and 9.
7. The seventh chapter will give research results on an approach that is based on decomposition of the assessment of the occurrence risk. Unfortunately this way of decomposition did not improve validity.
8. The eighth chapter will give the results of a replication of the validation of the assessment of the occurrence risk as it was done in the framework of the audit risk model. Four governmental organisations will be investigated. In this part of the study risk assessment only shows very weak correlations with the error rate or other validation criteria.
9. The ninth chapter will give the results of a study of the quality of tests of controls and errors of the previous year as a validation of risk assessment. It does so by investigating their predictive power for the error rate. As for the occurrence of errors, tests of controls appear to be predictive, but as for the size of errors no predictive quality is found.
10. The tenth chapter will discuss the results of this study. It will end with a suggestion to continually validate risk assessment and next to that also base the size of a statistical sample on the distribution of the error rate conditional on the assessed risk.

Chapter 2: Validation of Risks

Risk or probability - the 'real risk' - can only be observed by the varying amounts or numbers, resulting from processes, governed by this risk or probability. Our intuition tells that the error rate is such a number, but what is its quality as a validation criterion for the assessment of OR; how close does it come to the 'real risk'? Via a metaphor of rain, rivers and dykes, representing processes resulting in heights of water in a river (parallel with error rate), flooded dikes (excess of materiality) we try to justify the four validation criteria for the assessed risk and we get more insight in the 'real risk'.

2.1 Introduction: rain and risks

We start this chapter with a short story on excessive rainfall, with high risks of a flood as a consequence, not only depending on the rainfall, but also on the water stowing capacity of the German and Dutch rivers and forelands. This story offers a metaphor for the introduction of the inherent risk and the internal control risk. We elaborate this metaphor in section 2.2, with ideas for validation criteria for risk analysis as a result. We go deeper into these validation criteria in section 2.3. In section 2.4 we will take quite a different point of view on validation, by wondering whether tests of control can serve as a validation for risk assessment.

The story on excessive rainfall is about risk management in an extreme form, because the possible consequences of mismanagement are tremendous. In such a context, not only the metaphor as to risk management is interesting, but the question how to validate the assessed risks is even more fascinating, especially in the context of this thesis.

In the daily "Trouw" of Saturday December 11th 2004, an article "Niet ophogen die dijken, maar verleggen" ("Do not raise those dikes, move them") discussed the necessity of increasing the height of the dikes along the rivers in the Dutch estuary. This 'raising' was thought necessary after the near crises of 1993 and 1995. In 1993 the waterflow was of an extraordinary volume, causing an extremely high water level. In 1995 it was even higher, which made the authorities decide to evacuate many tens of thousands of people from areas next to the rivers Waal and Maas. The authorities thought this evacuation necessary, because the threat of a flood was thought to be very high. These water levels were the consequence of extremely long and intensive rainfall, causing a flow of water of unknown volume.

The article discussed the plans to create areas that could serve as a reservoir for the surplus of water streaming down the river in case its level would be so high as to cause too large a risk of flooding areas that – different from the forelands - are not prepared to suffer a flood. This reservoir would be planned to be given up to the river water, by deliberately opening flood channels from the rivers into these reservoirs, thereby keeping the water level under a critical level. Even so, there are villages in these so called "mega-bathtubs" and, to save them from being flooded, they would be protected by a ring-dike with a height of 6 meters.

Activists opposing these plans, argue that the worst case scenario, assuming that 18.000 cubic meters of water (per second) could stream into the Netherlands at Lobith, is too pessimistic. In Germany the dikes are being moved instead of raised, giving

more room for the water in the German forelands. Therefore a scenario in which 16.500 cubic meters per second stream into the Netherlands is more realistic as a worst case scenario. When a policy is chosen in the Netherlands, similar to that of the Germans, with the motto: “give room to the river”, the building of ring-dikes around villages and towns can be avoided.

So far for the short story

2.2 Risk of a flood: how to validate?

In this section we elaborate the "high water metaphor", by observing some typical traits in the story regarding risks and their control and by observing how assessments of critical probabilities are validated in the real situation to which in this story applies.

2.2.1 Risks

The first observation is that we are talking of “risks of a flood". We are talking about risks, a (small) chance of an event happening. We are talking of risk and not merely of chance or probability, because the consequences of the event happening are very serious: the impact of a flood will be tremendous. This is the reason why in the literature (see for instance Keeney and Raiffa, 1976, French,1988) risk is often treated as an expected loss: the risk of an action associated with losing €100,000 when the probability of losing it is one percent, is equal to €1000. So risk is associated with probabilities of an event and with (negative) consequences of that event.

In a second meaning risk merely is a probability or a chance of some (negative) event happening, for instance "the risk that it will rain tomorrow". Here the possible loss is not included in the concept, so a "high risk" of something happening, is "high" regardless of what is at stake, except for the fact that the concept of risk is used because negative events are involved.

The concept of risk in the audit risk model has the second meaning: it stands for "the risk of not detecting a material error in an annual account" if indeed there is one in it. “Risk” in practice is handled as a probability and it only refers to an expected loss as far as the event possibly happening regards something undesirable: the existence of a material error. This risk is also referred to as 'the risk of the auditor', and thus has the meaning of the probability of overlooking a material error. And although the consequences of such an event may be very large, these consequences are not included in the regular way auditors deal with this risk in the audit risk model. (See Klijnsmit et al., 2003).

Only recently proposals appear as to include expected loss in an adapted audit risk model. This especially so, in the light of the risks that audit firms appear to run of even going bankrupt, when they are involved in accounting scandals like that with Enron, WorldCom and Ahold (see again Klijnsmit et. al., 2003, Mollema, 2003, 2004). Whether these suggestions will be acted upon or not, confusion as to the meaning of risks lies in wait.

In this thesis we will adopt the meaning as it is used in auditing: risk is the probability of overlooking a material error. This is the logical consequence of the aim of this thesis: to validate the risk which the auditor assesses in his risk analysis.

2.2.2 Context

The second observation in the "high water metaphor", is that there are risks in a context:

- the way the landscape is formed, environmental conditions like the presence of forests and other conditions affecting its water retaining or absorbing capacity
- short-term and mid-term weather conditions like: will it rain or not, will it rain day after day or with larger intervals
- the question whether the melting water from the glaciers will grow, due to the rise of temperature of the atmosphere.

The whole of these (and other) risks due to conditions is analysed and aggregated into the probability formulated in the next question: “What is the probability of a period of such intensive rainfall that the flow of water will (almost) exceed the stowing capacity of the river?” This contextual risk (or probability) in the context of this “risk of a flood” cannot be influenced, which is a reason to treat it as context.
This does not mean that human actions do not affect these risks: the ruining of the woods, the canalization of the rivers and many other human actions, all affect the water retaining, absorbing and stowage capacity. Next to that, global warming affects the climate, rainfall and melting of the glaciers.

Risk analysis in auditing also deals with risks of the context: things that are not controllable with the given accounting methods, like competition with other organisations, nature of the organisation, nature of the product, attitude of the personnel, robustness of production methods, etc.. We have seen in 1.2.1, that these are called ‘inherent risk’.

2.2.3 Control

The third observation is that there are also attempts to control the risks of a flood, as they result from the risks of the context or from the "environmental risk".
One way of controlling the risks is by keeping the probability of a flood under control:
- dikes are raised (if necessary relatively instantaneously by means of sandbags),
- flood-control dams are built and once built, opened (and closed) when necessary,
- ships are urged to lower their speed or are prohibited to sail.

It can be argued that the probability of a flood is also controlled by adopting a proper organisation, for instance, with a proper separation of duties like: ship owners should not be made responsible for the closure of a waterway.

A second way to control the risks of a flood is to reduce the damage that may be the consequence of a flood. In the estuary there are not many possibilities to do so. Near the former Zuiderzee houses on the isle of Marken were built on posts. To a certain extent the situation with forelands is a way of damage-control: the forelands are used in a way as to only suffer to a minimal extent from being flooded. This use sometimes extends to the building of houses in the forelands on posts (or on a similar construction) or on a mound. The plans mentioned in section 2.1 to create bathtubs, also aim at damage control.

Risk analysis in auditing also deals with risks (of a material error) that can be controlled. More precisely: the risks are controlled (or hoped to be controlled) by a proper organisation of the administrative processes and more in general by a proper organisation of the business processes. Risk analysis assesses the extent to which these controls will be, or may have been successful. Next to this in auditing the probability is kept under control of not seeing an error that could occur due to the

imperfect control of the causes for an error. This is realised by applying substantive procedures to a sufficient extent.

Damage control can also be found in auditing. Some errors (for instance 'just some mistake in a VAT-calculation causing an error of €100, -') are deemed to be less serious than others (for instance some fraudulent appropriation of money, causing an error of €100,-). 'Damage control' consists of adapting the level of materiality to the kind of error. For serious errors, for instance errors that could have a significant political impact, a very low level of materiality is adopted. Of course it must be hoped and it may be expected that the organisation itself also deals with 'damage control'.

2.2.4 Materiality

A fourth observation reveals the key issue of it all: flood relates to a critical level of the water. Thus, a flood of the forelands is not critical; an auditor might say that flood of the forelands is 'not material'; it is an 'error' which does not influence decisions as to the use of land on the land side of the winter-dike. But once the level of the river gets too high, there will be a flood: the water will go over the winter-dike and flood the hinterland. A water level that is equal to the height of this dike is "material". This is like the critical "level of an error": this is called material when it is higher than the level of materiality. In our metaphor 'flood' stands for the 'necessity of a qualified opinion'. So the similarity can be extended: a higher dike means a smaller probability of flood of the hinterland; a higher level of materiality means a smaller probability of the necessity of a qualified opinion.

There is a basic dissimilarity included in this part of the metaphor: a higher dike means more expenses and more safety, whereas a higher materiality means less expenses and less safety, because a less intensive audit is needed to realise an audit risk that is acceptable. Of course users of the financial statements may consider such a higher level of materiality undesirable

So far the observations make explicit how the "high water metaphor" provides an image of the accounting and audit process, in particular regarding the concepts of risk analysis. We now wonder whether the metaphor can be extended to the validation of risk assessment. To this end we continue with two observations, the first in 2.2.5 and the second in 2.2.6.

2.2.5 Water authorities and validation

The first of these continuing observations results from an extension of the high water metaphor. The extension regards the so-called "Waterschappen", Dutch for "Water authorities". The following can be said about "Water authorities".

In the Netherlands we have authorities that are responsible for the safety of the land against floods. This is embedded in an age-old tradition. In the last decades their responsibilities have been extended to the quality of the water in - what in the Netherlands are called - the "bosom", or the "bosom waters", the system of canals and lakes, ditches, brooks, streams, channels, trenches and rivers and any other sort of condition in the landscape that can hold and store water. The "bosom-water" is of paramount importance in the Netherlands in dealing with precipitation for many purposes: keeping it in times of shortage of water and when precipitation is too abundant, guiding it as quickly as possible to places where it will do no harm: the IJsselmeer, or even better the North Sea.

The authorities that deal with this water management have the legal duty to build and maintain dikes that safeguard the land against floods. Therefore, among many other things, they have to establish the height of dikes that is necessary to be safe against floods. Until 1995, this especially was thought to be a very relevant problem for the parts of the Netherlands which are in direct, open connection with the North Sea. Since 1995 we know that also upstream this is an important problem.

These "Waterschappen" aim at a rate of security, which is such that with a very high probability no floods will take place. The norm is that only with the exception of - on the average – say once in 1000 years a flood might take place. In order to be able to comply with this demand the "Waterschappen" need a fair estimate of the probability of this event, so of the probability of the water flooding over a dike of a certain height. Therefore they analyse all water heights during the last decades or hundreds of years. These heights form a distribution to which some probability model can be fitted. This model can be transformed into another model, which can be assumed to be valid for a period of 1000 years. The upper-tail probability in this distribution of the dike height - the probability that a water height will exceed the height of the dike - is the probability which is needed (see Groeneboom (1992), Wijbenga et al (1993), who take once in 1250 years). It may be seen as the validation of the assumed "once in 1000 years".

This approach will turn out to be applicable to our validation problem: we could subdivide our cases after the assessed risk, so that we get four subclasses: one for risk ‘very low’, one for ‘low’, one for ‘medium’ and one for ‘high’. For each class we form the distribution of all observed error rates, the empirical distribution and we look for a theoretical distribution that fits this. For the materiality in a specific case we can calculate the probability that an error rate in this theoretical distribution might exceed this level of materiality. This upper-tail probability corresponds to the assessed occurrence risk. It will vary with the location of the theoretical distribution and its dispersion. For low risks, the location of the distribution may be expected in the region of smaller error rates, for higher risks in the region of higher error rates. When at the same time the dispersion is approximately equal for the various levels, or increases for the higher levels of occurrence risk, it can be said that a shift in the location of the theoretical distributions to the higher error rates should correspond to higher assessments of the occurrence risk, in case of a valid assessment.

Figures 2.1 and 2.2 illustrate the idea of how a theoretical distribution, fitted at the empirical distribution, may look for an assessed occurrence risk OR at ‘low’ and how it may look for OR=’high’. The shaded areas depict the probability that an error may exceed the materiality. It is obvious that the two figures are based on an assumption of valid risk assessment.

Figure 2.1: empirical risk for OR='low'

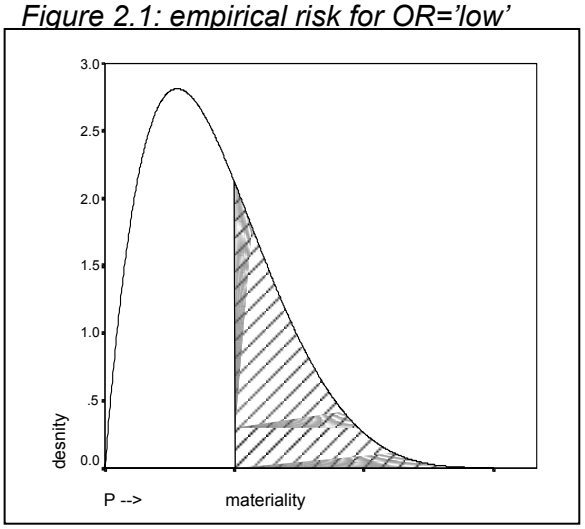
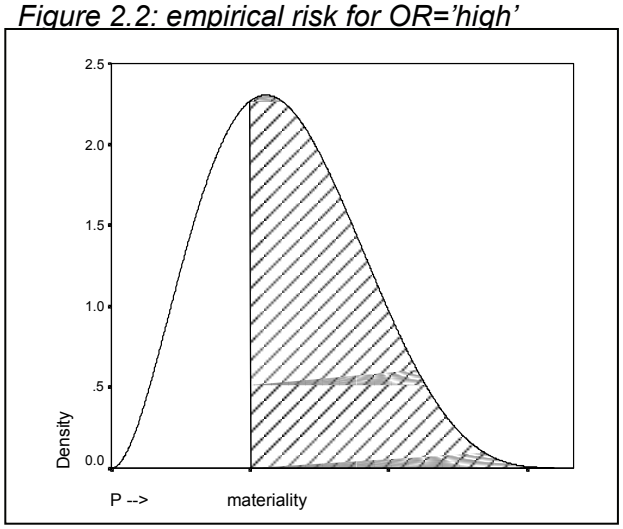


Figure 2.2: empirical risk for OR='high'



2.2.6 Weather forecasts and validation[6]

The second of these continuing observations regards the way weather forecasts are validated.

In weather forecasting the weather is seen as a process, the outcome of which is the particular weather at a certain moment, or day. This process is modelled in a very precise way by means of relevant physical laws, such as the law of Boyle-Gay Lussac, the law of Buys Ballot, laws of thermodynamics, etc.. The model is fully deterministic and meteorologists have succeeded in making it of a high quality. This can be deduced among other things from the fact that its predictive power is largely enhanced by feeding it with data of a resolution of a measurement in every 20 by 20 km instead of in every 50 by 50 km. These measurements are done worldwide and the weather predicted also is worldwide. A deviation of the 'real weather' from the predicted weather is attributed to measurement imprecision.

In order to deal with the more or less random character of the weather, small (random) disturbances are entered into the model, every disturbance possibly leading to some other (type of) weather, provided disturbances are chosen that are not deadened in and by the local circumstances. These disturbances can be seen as representing the possible imprecision of measurement. Part of the professionalism of the meteorologists is to choose the appropriate disturbances at the appropriate places. In this way, in practice 50 different "weathers" are calculated: for every 20 by 20 km (once) or 50 by 50 km (49 times) point on earth the expected temperature, air pressure, precipitation, wind direction and speed etc are calculated. What reality can not do: letting 50 different results at the same time come from one process, is realised by the model. From these 50 different results probabilities can be deduced. If for instance in 10 of the 50 "weathers" the amount of precipitation exceeds 10 mm, the probability of at least 10 mm of precipitation is assessed to be 1/5.

This methodology of weather forecasting is known as the "Ensemble Prediction System", EPS (see Floor, 2002) and a validation study, using the approach outlined above, can be found in Kok (2001). Validation of the probabilities in the EPS is very straightforward, because the way a probability is assessed in the forecast is very close

[6] I owe a great deal to Mr Robert Mureau of the Royal Netherlands Meteorological Institute (KNMI), who introduced me to the Ensemble Prediction System (EPS) of weather forecasting and the way probabilities are calibrated in this methodology.

to the (Laplace) definition of probability: the rate of relevant events in the set of all possible events (see Hogg and Craig, 1970, p11). Actually, validation of the assessed probability in the precipitation example above is done by looking at the relative frequency of precipitation of at least 10 mm in all the "weathers" in which this event was forecasted with the same probability of 20%. Every other probability can be validated in a similar way.

It is evident that well validated weather forecasts are of paramount importance in situations like the crises sketched in section 2.1. Here, the decision to evacuate people or not, heavily depended (among other things) on weather forecasts for the short-term: the precipitation to be expected in the first couple of days, the expected wind direction and force.

2.2.6.1 Risk assessment as a mental sensitivity analysis

Now, just like the weather, the administrative processes only produce one reality. But, unfortunately, in auditing the administrative processes are hard to simulate. For this to work, some algorithm would be necessary in which for varying values of the determinants of the error rate the resulting error rate would be calculated. Evidently, determinants in this case are the properties of the environment, the context of the business, the relevant controls and other properties of the administrative organisation. So far such algorithms do not exist (to our knowledge). But this will not prevent the auditor doing some mental, professional, guessing of the form of "what-if …": what if all separations of duties work perfectly?, what if the separations of duties work just a little less perfectly?, what if access control in the EDP system is poorly organised?, etc.. These questions will form a kind of mental sensitivity analysis, with an eye on the error rate in the account under audit that results from the administrative process.

The result of these what-if questions can be imagined to be a set of possible error rates in the account under audit. Each error rate will be associated with its own likelihood. There will be a most likely error (not to be mixed up with the MLE, the statistical estimate of the error in the account, to be discussed in 2.2.6.2, 3rd point) error rates close to this will also have a relatively high likelihood and there will be less likely error rates. In general the expected likelihood will decrease with the distance of the associated possible error from the most likely error. The portion exceeding the level of materiality of these possible error rates and their accumulated likelihood represents the occurrence risk, OR. The auditor will associate risky processes with expectations of relatively high errors and thus, dependent on the level of materiality, with a high OR. And for lower assessments of the risk he will have the lower levels of possible error in mind and therefore the lower levels of OR. This view on risk assessment can be depicted in a figure almost the same as the figures 2.1 and 2.2. The only difference would be that the p-axis is not based on the empirically found error rates in a set of similar accounts, but represents the p's (error rates) that are deemed to be possible according to the risk assessment in one account under audit. The shaded area in the pictures represents the occurrence risk. We could say, in Bayesian terms: the figures represent the prior distribution of p (formally: f(p)) that follows from the risk assessment by the auditor.

2.2.6.2 Four possible ways of validating the assessment of OR

This mental image of risk assessment offers a perspective on four possible ways of validating the assessment of OR.

For the first two ways we observe that the impossibility of letting one process produce several outcomes (the strict parallel with the EPS), can be circumvented:

(1) by clustering processes that were assessed to have the same occurrence risk. The distribution of errors found in cases with OR='low' should tend to show

smaller errors, especially for the same level of materiality, than the cases with OR='medium' and even stronger for the cases with OR='high'. This validation is parallel to that at the end of the previous subsection. By fitting a probability model to this empirical distribution (as is done in Groeneboom, 1993) the empirical distribution is smoothed and calculations can be made easier; also some probability can be given to values larger than the largest observed, which in a probabilistic sense must be deemed to be realistic. (Compare this to modelling observed lengths in a sample of inhabitants of a country by means of a normal distribution; in this distribution there will always be a positive probability of lengths larger than the largest observed).
This results in a first validation criterion: *the location of the conditional empirical distribution of error rates*

(2) by observing that in the Bayesian approach a probability model on possible values for p (the error rate in the population), given the data (D) (in one audit case) is given by

$$f(p|D) = (f(D|p)f(p))/f(D) \quad (1)$$

So the probability of a value of p (formally: the probability density function of p) depends on the likelihood of the data, given this value for p (f(D|p) and on the prior distribution of p (f(p))[7]. When we take f(p) to be non-informative, f(p|D) virtually only depends on the data in the case we analyse. This means that risk assessment can be validated by the information from one case, by calculating the posterior distribution of p, given the data D and with a non-informative prior. In this posterior the probability of the values for p larger than materiality is the risk of a material error (given the data). We again can refer to figures 2.1 and 2.2, where now the curves represent the posterior density of p, given the data (from one case) and where consequently the shaded areas represent the risk of a material error, now given these data. In formula this shaded area gives:

$$P(p>M|D) \quad (2)$$

Actually, when a non-informative prior is being used, (2) gives the sampling risk, when D are collected by means of a sample.
This results in a second validation criterion: *the sampling risk*

(3) For the third way we observe that in the line of reasoning of this subsection processes of high quality will coincide with no or only small error(s), so the error in the account itself is a good indicator of the quality of the administrative processes. In practice we will not know this error, but we have an estimate, resulting from substantive testing possibly completed with other substantive procedures. In the practice of auditing this estimate often is called the most likely error (MLE). Note that this is a statistic, resulting from classical statistical procedures, so basically different from the result of the 'mental sensitivity analysis' discussed in 2.2.6.1, which also lead to a 'most likely error'.
This results in a third validation criterion: *the most likely error, MLE*

(4) For the fourth way we look again at the weather forecasting metaphor. We observe that often a weather forecast has the form of: "tomorrow it will rain" or "tomorrow it will not rain". The most simple validation of this forecast is to establish whether it rained next day. The event of rain can also be used to validate the probability forecast, because it may be expected that days with rain are preceded by weather forecasts with a higher probability of rain than days where it did not rain. The logical parallel of this simple validation is to validate the occurrence risk (OR) with the actual occurrence of a material error. It may be expected that in cases with a low OR excess of materiality will occur less frequently than in cases with a high OR. In this thesis, as a rule only results

[7] The denominator, f(D) is constant for all possible p's; it only serves as a scaling factor, leading to a correct summing up of the probabilities to 1.

on sub-accounts of the annual accounts were used, so we will not be in the position to make use of the actual audit opinion. But a similar validation criterion can be found in what we will call the "audit position". With 'audit position' we mean the place of the most likely error estimated from the audit relative to the materiality level. This 'position' is unqualified or "OK" if the most likely error is smaller than the materiality.
This results in a fourth possible validation criterion: *the 'audit position'*.

2.2.7 "Real risk" revisited

Our discussion in 2.2.6 gives a possibility to be more precise about the concept of "real risk" as we used it in the generic question for this study in section 1.1. In auditing, the risk of a material error is associated with the annual accounts. But this gives definition problems:

- for the possibility of a material error in given annual accounts, in classical statistics the concept of probability, is not defined, because in this approach the given accounts either contain a material error, or they do not. Probability (and consequently: risk) only gets a meaning when some random selection is in operation, of which the outcome has to be predicted. But once the accounts are there (possibly errors included), there is no random process influencing existence or size of the error. So from the viewpoint of classical statistics, the concept of risk does not apply;
- in Bayesian statistics, it is possible to define a prior distribution on the error rate and let the upper tail probability of materiality in this distribution be the occurrence risk; we discussed this possibility in 2.2.6. But classical statisticians have a serious objection against this approach: they stress that a form of random selection is needed, or else the concept of probability makes no sense. And in the prior distribution this randomness is missing. Moreover, prior distributions often express a subjective opinion, causing the same situation to give rise to different priors, for varying subjects.

Some reconciliation of both approaches is possible by making explicit that the distribution of possible error rates is directly linked to the administrative processes and the controls included. The auditor then does not assess the probability that the annual accounts contain a material error, but assesses the probabilities associated with the possible outcomes (error rates) of the underlying administrative processes. This assessment produces a legitimate probability distribution in the classical approach. And then classical and Bayesian approach are consistent; the Bayesian may say that the prior he wants to establish refers to the possible error rates and their associated probabilities to be produced by the process.

This distribution in principle does not differ from classical ones like the distribution to be expected when throwing a dice ten times and establish the distribution of the sum of the outcomes for each 10 outcomes.

It leads to the conclusion that 'real risk',

- in case of the occurrence risk, is the risk that the administrative processes produce annual accounts with a material error,
- in case of the audit risk, is the risk that a material error is overlooked, as a result of occurrence risk and detection risk.

2.3 Validation criteria

In this thesis the four entities introduced in 2.2.6 will serve as validation criteria. We summarise them in a different order as follows:
1. The error rate or most likely error;
2. The 'audit position' of the error rate;
3. The P(p>M|D), the sampling risk;
4. The distribution of the error rates conditional on the risk assessment

We will give them a closer examination.

2.3.1 The error rate[8]

In section 2.2.6 we concluded that the error rate in the annual accounts is an indicator of the quality of the administrative processes that produced these accounts. We also argued that there is a random part in the outcome of these processes and therefore the possible error rates in the account under audit may vary although the quality of the process is constant. But in practice this variation will be very small, because of the large number of transactions. So the error rate in the account almost perfectly indicates the quality of the process.

In most situations we will not know the real error rate in the annual accounts or sub-account we are auditing (if we would, we would not run an audit risk). We only have estimates at our disposal, based on some substantive procedures, resulting in, in auditors' terms, the most likely error, MLE. In cases where the audit risk is assessed at low, the MLE may be expected to be low. In cases where the audit risk is assessed to be high, the MLE may be expected to be high. So in general a valid risk assessment will lead to positive correlation between assessed risk and error rate. But two aspects still have to be taken into consideration.

The first one is that an error rate of some size will indicate more audit risk when the level of materiality is smaller. So only at a constant level of materiality the same error rate will correspond to the same risk to be assessed. The correspondence between error rate and assessed risk may be expected to be stronger at a constant level of materiality.

The second aspect regards something similar with regard to the sample size. This aspect is relevant in cases where a sample is used to estimate the error in the sub-account, which will be the normal situation in practice. In such situations the same distance between error rate and materiality indicates more audit risk when the sample size is small than when the sample size is large. Because, with a large sample size, there is more reliability in the statement that materiality will not be exceeded in the sub-account than in cases where the sample size is small. So the correspondence between error rate and risk assessment holds at its strongest within subgroups of cases with the same level of materiality and sample size.

In this study we did not get enough data to make the corresponding subdivisions in which to establish the relation between error rate and assessed risk.

[8] In this thesis 'error rate' represents the relative *size* of the error, so size of error divided by size of account. When we mean the occurrence of errors, regardless their size, we will use 'occurrence' or 'fraction of' or in some other way let the context make it clear.

2.3.2 The ‘audit position’.

We discussed the possibility to validate the assessment of the occurrence risk with the rate of annual accounts in which the estimated misstatement is larger than the level of materiality, found from a set of annual accounts produced by administrative procedures which were assessed to have the same OR.

By validating risk assessment directly for cases with the same occurrence risk, we could ask the question whether similar administrative processes lead to a similar assessment of the occurrence risk. But in this study we will not have enough cases to give an answer to that question, because there are many more kinds of administrative processes than there are levels of assessment (at most four). In the end this is no problem for our validation because, as for this "similarity", it may be assumed that, even if in practice businesses and types of processes may differ considerably with regard to their procedures, risk assessment answers as precisely as possible the question to which extent the inherent and control risks are covered by the set of controls. So in this study only the level of the assessed risk is relevant for the similarity, which means that in principle procedures with the same assessment of OR can be seen as "similar".

If the error rate for the complete annual accounts is larger than materiality, the audit opinion will be qualified. In 2.2.6, for sub-accounts we introduced the “audit position”, the definition of which can be made more precise as follows:

Definition of ‘audit position’:
The ‘audit position’ of the error rate is its position relative to the materiality; it is “not OK”, if it is equal to or larger than materiality; it is “OK” if the error rate is smaller than materiality.

In practice three or four levels of inherent -, or control risk, or of the occurrence risk are in use: “very low”, “low”, “medium”, “high”. Now each of these risk levels could be validated by looking at the observed rate of “not OK positions" at that level. It would be hard to decide on an exact rate that should correspond to this level, because this could at its best only be given by agreement. But it may be expected that with an occurrence risk assessed at low there will be relatively much “OK positions” for the error rate in the corresponding set of sub-accounts, whereas with an occurrence risk assessed at high there will be more “not OK positions”. So a valid risk assessment will lead to a positive correlation between assessed risk and the rate of “not OK positions” (the higher the assessed risk the more “not OK positions”).

It should be noticed that taking the error rate as decisive for the ‘audit position’ will mean that in this study we will use the most likely error (MLE), as estimated by the auditor, as a criterion for the ‘audit position’ being “OK” or not. In practice often the upper error limit (UEL) will be used as the criterion for the decision on the audit opinion, in order to realize enough reliability with regard to this audit opinion. This UEL is a stricter criterion (for the audit opinion) than the MLE (for the ‘audit position’); there will be more qualified opinions than “not OK positions”. But it does not mean that the ‘audit position’ of the MLE cannot be used as a validation criterion.

2.3.3 The sampling risk

In 2.2.6 we concluded that P(p>M|D) is a natural indicator for the quality of the administrative processes that produce the account under audit. This probability can be expressed with greater precision as P(p>M|MLE), the probability that the error rate p in the population will exceed materiality, given the actual estimate of this real error.

In chapter 4 we will show how to calculate the this probability: P(p>M|MLE), (1)

This use of the sampling risk is further justified, because it solves the two problems we mentioned in 2.3.1 (dependence on materiality and dependence on sample size), regarding the use of the error rate as a validation criterion. Because when we use the sampling risk, sample size and distance between estimated error and materiality, are taken into account. So the error rate as a validation criterion is in a sense corrected for its distance to the materiality and the size of the sample in which it is observed.

For a second further justification we refer to our discussion in 2.2.6 that the occurrence risk can be seen as the probability of all error rates which are higher than materiality and that it is this probability that we estimate with probability (2) in 2.2.6.2.

The sampling risk has one major drawback: it is dependent on the sample size, and one major positive property: it is calculable in a single case. In this single case, the level of materiality co-determines the validation. In our data analysis we will compensate for the drawback by also using a standardised SR, by imputing constant sample sizes. We will explain how, when we do the analysis.

A second consideration regards the dependency of the sample size on OR: the auditor plans the detection risk dependent on the assessed occurrence risk. This results in allowing a relatively high DR (and consequently a relatively small sample size), when OR is assessed at ‘low’ and also planning a relatively low DR (and consequently a relatively large sample size) when OR is assessed at ‘high’. But considering that in practice the errors may vary only in a small range, a sample with small size will tend to result in a relatively high SR and a sample with a large size will tend to result in a low SR. This consideration results in an expectation that low OR coincides with high SR (given the sample results) and high OR coincides with low SR (also given the sample results), so a negative correlation is to be expected, according to this logic. Now, the logic is far from absolute: it is not necessary that low (high) OR always coincides with a small (large) sample size and the error rate may grow with the assessed OR. But the dependency of SR on the sample size certainly is a disturbing factor. It is an extra reason for using SR with a uniform sample size, as discussed at the end of the previous paragraph. We also will calculate the correlation between sample size and OR. When this appears to be small, use of SR as a validation criterion remains appropriate.

Because SR has the dimension of a risk (a probability) and because it is based on all relevant variables: error, materiality and sample size, it may be expected to be the best validation criterion. The caveat we have to make for the dependency of the sample size on OR, makes the standardised SR (with the sample size standardised) also a candidate for the best validation criterion.

2.3.4 The empirical distribution of the error rate

In section 2.2.5, we introduced the empirical distribution of the error rate and the possibility to fit a theoretical distribution to it. The location of this distribution can be used as a validation criterion for the assessment of the occurrence risk.

For a given assessment of the occurrence risk, say "low", the error rates can be collected from all audits where the occurrence risk was assessed at "low". These error rates (the water heights) form a distribution in which the probability that the error exceeds some level of materiality (the height of the dike) is the occurrence risk that should be associated with the assessment of "low" at the given level of materiality. When the assessment is valid, this upper-tail probability may be expected to be small,

due to relatively many small error rates. The upper-tail probability of the materiality increases with an upward shift of the distribution. So for cases with assessment "medium" or "high" we may expect a shift of the distribution to the higher values of the error rate.

The empirical distribution has one major drawback: it cannot be calculated in a single case and related to this: it cannot fully take into account that a single risk assessment is dependent on the level of materiality. This would ask for an analysis of the empirical distribution for each distinct level of materiality and so for many more cases than available in our study. Next to this limitation, it has at least two positive properties: it is (relatively) independent of the sample sizes and, related to that, it gives a relatively easy possibility for calibration of the assessment of OR. Our research will not go as far as calibration of the assessed risks, but it is worthwhile to indicate the possibility, which we will do in the next subsection

2.3.5 Validation and calibration

In short, 'validation' of a variable can be seen as answering the question: "Does this variable, given the way it is measured, measure or indicate what it is intended to measure or indicate?" Yet another question is: "Does this variable, given the way it is measured, give the right value for what it is to measure?". This latter question refers to 'calibration'. For risk assessment validation asks for measures that vary with the assessed risk (when this assessment is valid), but these measures need not give the 'right' value for the risk to be assessed. Would we wish to calibrate risk assessment, it would ask for some unambiguous quantification of the assessed risk The error rate does not have the dimension of a risk, the sampling risk is dependent on the sample size and the 'audit position' in principle could serve as standard, but it is crude: it either is 'OK' or 'not OK'.
But one could form the empirical distribution of error rates from a set of audits, grouping them after level of materiality and assessed OR. Validating risk assessment by means of the upper-tail probability for each of the groups, in the empirical distribution as done above, is not only fit for validating but also for calibrating the assessed risks. It directly gives an upper-tail probability that can be connected to the assessed level of OR. In this thesis the distribution of the error rates will be modelled by the beta distribution; this distribution is appropriate, as the error rates form outcomes between 0 and 1 (0 and 1 included) and the beta distribution is very flexible: it can take forms from symmetrical to highly asymmetrical .

It is clear that the suggested way of calibration asks for a large number of case per group. In our research we did not have enough cases and therefore we do not try to calibrate the assessment of OR.

2.4 Validation by tests of control?

So far we have discussed four possible criteria for validation of risk assessment. They are all related to the error rate. In this section we introduce quite a different perspective on this validation, which is implicitly taken on many occasions of risk analyses in practice. It is the perspective in which tests of control are implicitly seen as a means for validation. This follows from what is stated in ISA 200 paragraph 21: "When the auditors assessment of the risk of material misstatement includes an expectation of the operating effectiveness of controls, the auditor performs tests of controls to support the risk assessment." It also follows from similar prescriptions in HCDAD (1997) and other manuals, which have as a result that reduction in the extent of substantive testing only

is allowed when the risk assessment on which this reduction is based, is supported by tests of control. However, neither ISA 200, nor ISA 400 (par. 24-34), nor HCDAD (1997) make explicit how tests of control add to (or validate) the quality of risk assessment.

We will not directly try to make the contribution of tests of control more explicit, but we will test a necessary condition for that contribution to be valid. We will investigate whether tests of controls have predictive power for the error rate, should they serve as support for an assessment of control risk at less than "high". This condition is necessary because supporting information that does not predict the error rate to a certain extent, cannot serve as a replacement for direct substantive testing. And consequently it can not serve as a support for risk assessment as far as it is used as a replacement for direct substantive testing.

Chapter 3: Audit Opinion: Judgment under Uncertainty

Much research is done into the quality of risk assessment. The classical paradigm for this research is that developed by Tversky and Kahneman. Many of our references regard research in the tradition of this paradigm. They investigate a variety of ways in which information can be presented and how mental processes deal with it. Other references use relative quality criteria like consistency. We will give all references a treatment in their own right, in trying to find answers to questions raised by risk analysis on an audit object.
Only a few studies validate the assessed risk on the error rate (or some variable based on it). Where in our research this is a key criterion, we had to construct a bridge between the available literature and our aim to validate OR on the error rate (or related variables). This aim is decomposed into four questions that will be leading for our research.

3.1 Introduction

Issuing an audit opinion is being engaged in judgment under uncertainty. Judgment under uncertainty is the label under which Amos Tversky and Daniel Kahneman (1971, 1974, 1981, see also Kahneman 2003) did their pioneering research in the 70's of the last century into the question how human beings deal with uncertain or incomplete information on which they have to base a decision. In this research they focused on the heuristics people use to deal with this incomplete or uncertain information. By "heuristics" they mean a rule or a set of rules that govern the way information is processed by the human mind. Every day examples of these heuristics can be found in sayings like "one swallow does not make a summer", or "things are never as bad as they look", or "the better is the enemy of the good"[9]. These are all heuristics, or heuristic principles, which give guidance to the human mind when deciding on a course of action or when giving predictions, based on incomplete or uncertain information. Many of the things Tversky and Kahneman found apply to the judgment that leads to an audit opinion, especially when risk analysis contributes to this judgment.

We will continue this section with a short overview of the four articles mentioned above, illustrated with examples from auditing and some other fields. In section 3.2 we will speculate on the question how the heuristics discussed in the first section might occur in auditing. In the 3rd section we will discuss the validity of the audit risk model. In the 4th section we will discuss relevant literature dealing with heuristics and related approaches in audit research. In the 5th section we will discuss the results of our survey of literature. In the 6th we will connect our research questions to these rsults. We conclude with a table in which the results of the articles discussed is summarised.

3.1.1 The Law of Small Numbers

Tversky and Kahneman (T&K, 1971) state that people believe in the "law of small numbers". This "law" entails that people regard a random sample from a population highly representative of this population, almost irrespective of the size of the sample. They overlook the fact that especially samples of small size are relatively likely to display atypical pictures of the population they are drawn from. This means that the law

.9 See Wagenaar (1977) for a more elaborate account on the way sayings and proverbs serve as daily life heuristics when people deal with incomplete or uncertain information.

of small numbers causes people to come to biased conclusions about the population, with a relatively large probability. One of the causes for this bias is the idea that randomness compensates for outliers. People believe that if in a random draw an extreme value would be selected, the randomness will take care of compensation in one of the next draws. But randomness has no memory. It is not difficult to point at instances in daily practice of auditing, where auditors implicitly or maybe even explicitly capitalise on this law of small numbers. An example illustrates this.

Example
An auditor draws a random *MU-sample* (a sample of monetary units) of size 30 from an account in which 2% of the monetary units is in error. He finds no errors in his sample. Using this as the only information on this account, he concludes that the account will be error-free or at least well below his materiality level of 1%.
By doing this he disregards the fact that his decision procedure has a probability of 55% to find no errors. This probability was calculated with the help of the binomial law; it is the probability to find 0 errors in a sample of size 30, when the probability of 'success', p=.02.

In fact auditors capitalise on the "Law of Small Numbers" every time they decide to base their audit opinion on the results of a small sample, especially when they base their opinion on the point estimate, the most likely error, and not on some confidence upper limit for the estimation.

3.1.2 Heuristics and Biases

In their article in Science (idem 1974) T&K describe three heuristics:
(1) the representativeness heuristic, (2) the availability heuristic and (3) the adjustment and anchoring heuristic.

3.1.2.1 The representativeness heuristic
The *representativeness heuristic* can be seen as a generalisation of the belief in the law of small numbers. With it, T&K mean the phenomenon that people classify an object A as belonging to a class B, solely on the basis of the description of A and the extent to which it is representative of class B. In other words, A is classified as belonging to class B when it is typical of class B. T&K show that with this heuristic various biases are associated:

- the classification is insensitive to prior probabilities,
- it is insensitive to sample size
- it is insensitive to predictability
- it misjudges regression effects.

Overlooking prior probabilities, for instance, results in even probabilities when a non-informative description of a person is given as an indication of an answer to the question whether he is a farmer or a solicitor. In experiments the probability of the person to be a farmer is assessed at approximately 50 percent, even though the occurrence of farmers is much higher than of solicitors. So it would be more logical to assess the probability much higher than 50 percent.
In auditing this bias would occur when an auditor decides to give an unqualified opinion, based on a small sample, in an account of which previous audits indicate a high probability of a material error. This (hypothetical) course of action also suffers from the next insensitivity.

Insensitivity to sample size means that a point estimate derived from a small sample leads a person to the same conclusion as the same point estimate derived from a large sample. This is another way of stating the law of small numbers.

Insensitivity to predictability means that predictions of for instance the future value of stock or the demand for a commodity are based on a description of the company involved, without taking into consideration whether this description is really informative for what has to be predicted. So when a description of the company is very favourable, a high profit will be predicted, even if the properties that are described, are not relevant for the profit.
In auditing, this bias may occur when risk analysis is done on irrelevant properties. To speculate a little on (hopefully) a caricature: the financial director is looking very honest, most of the accounting staff are wearing ties, building and working environment are very well organised and brand new.

Misjudging regression effects is due to taking an extreme performance as representative. This causes a prediction being too high in most cases, because in general it is hard to perform the next time at the same level, or higher, when the last time was at the top of your ability. This effect is mirrored in the case of extremely bad performances: these will lead to predictions that are too low.
In auditing, this bias may occur when last year the auditee performed very badly. This surely will influence the assessment of risk for this year, regardless of present year's observations. (An effect that on other grounds can be deemed to be favourable).

3.1.2.2 The availability heuristic
The *availability heuristic* makes people assess the frequency of a class or the probability of an event dependent on the ease with which instances of occurrences come into mind.
Biases are associated with the following conditions:

- the retrievability of instances
- the effectiveness of a search set
- the imaginability of instances
- the illusory correlation effect.

The retrievability bias, for instance, occurs because famous persons are more easily remembered than less prominent persons, or larger line items have higher attention value than smaller line items. Selection bias due to this "retrievability by size of item" is shown by Hall et. al. (2001), who let subjects select items from an account, as random as possible. So not only judgment but also non-judgment can cause biases.

By a search set T&K mean a set of mental rules, or heuristics, that enables a person to retrieve information from his memory. 'Always be vigilant with lowly educated staff', 'always be vigilant with quantitative outcomes' may be such mental rules (implicitly guiding ideas), as part of a search set. Such a set may influence the choice of an audit object or influence the way information stored in memory, is processed.

Imaginability plays an important role in the evaluation of probabilities in real-life situations. The risk of engaging in some project often has to be assessed beforehand and then not only experience, but also imagination plays a role in coming at contingencies that may disturb the project. When this imagination is vivid, the risk may be overestimated because the likelihood of the contingency is much less than the vividness of the imagination.

The availability heuristic also provides a natural explanation for the illusory correlation effect. It occurs when two events occur simultaneously; human mind then assumes they are associated, correlated, or even related by some causality, even if there is no association at all. On future occasions, when one of the two events happens the heuristic makes the observator assume that the other will occur also. In the Netherlands an every day example of this "illusory correlation effect" is the remark often made in a company when a conversation suddenly falls silent: "The Vicar passes". The implicit assumption in this remark (meant as a joke) is that the company has reasons to keep quiet when a vicar is near, because of a bad conscience. (As a matter of fact it must be admitted that no formal record is known of instances where these two events co-occurred).

3.1.2.3 The anchoring and adjustment heuristic

The *anchoring and adjustment heuristic* regards the phenomenon that in many situations people make estimates by starting from an initial value, the anchor, that is adjusted to yield the final answer. Insufficient adjustment is synonymous with the anchoring effect. One of the instances in which it will occur is in incomplete computations, such as in solving the well-known 'joint birthday problem'. This goes as follows:

> *Joint birthday problem:*
> "*Give an estimate of the probability that of 23 people together at some party two or more have their birthday on the same date"*.

Most people intuitively assess this probability much smaller than the actual 50%. This is because they start with some small probability like '1 or 2 divided by 365' as an anchor and fail to adjust sufficiently for the numerous combinations in which the event for which the probability is to be estimated can occur.

Also adjustment biases appear in the evaluation of *conjunctive, disjunctive and simple events*. A conjunctive event is an event that, by definition, happens if (and only if) all of a set of composing sub-events happen simultaneously (are in conjunction); it is analogous to a series connection. A disjunctive event is an event that, by definition, happens if at least one of a set of composing sub-events happens; it is analogous to a parallel connection. A simple event is not structured as a combination of sub-events. Compared to a simple event, people tend to overestimate the probability of conjunctive events and to underestimate the probability of disjunctive events (which also happens in the birthday problem described above). T&K mention examples in which the equivalent formulations of a problem as a conjunctive or a disjunctive event had this impact on the assessment of the associated probability.

3.1.3 Framing

In their article in Science T&K (1981) introduce the concept of *framing* and the effects it has on decisions. Framing refers to the way a problem is stated. In daily language a message may be that the glass is half-empty, or that the glass is half-full. It is everyday experience that this difference in 'framing' of the same information, often causes differences in appreciation, or action.

By way of an example, T&K give the consequences of two programmes that have to cope with the threat of an outbreak of an unusual disease. The consequences were the same, but framed in a different way. They also give a sketch of an experiment in which these descriptions were given to experimental subjects that were asked to choose between the programmes. The following example (from T&K, 1981) summarises their experiment.

Example
Problem 1: Imagine that the U.S. are preparing for the outbreak of an unusual Asean disease, which is expected to kill 600 people. Two alternative programmes to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programmes are as follows:
If programme A is adopted, 200 people will be saved
If programme B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.
Which of the two programmes would you favour?
In an experiment this problem was administered to 152 people; the majority choice, 72% in this problem was risk averse and chose the option in which 200 people were saved: the prospect of certainly saving 200 lives is more attractive than a risky prospect of equal expected value, that is a one in three chance of saving 600 lives, chosen by the other 28%

Problem 2: The story about the disease is the same as in problem 1.
If programme C is adopted 400 people will die.
if programme D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.
In the same experiment this problem was administered to 155 other people; the majority choice was now risk taking: the certain death of 400 people , chosen by 22%, is less acceptable than the two in three chance that 600 will die, chosen by 78%.

T&K state that in general choice problems involving gains lead to risk averse choices and problems involving losses lead to risk taking choices as is illustrated in the "Asean disease" example.

Remark
Framing is different from the three heuristics in section 3.1.2: the heuristics can directly be seen as a kind of "mental aid" to solve a problem of information processing. Framing is a phenomenon that is inevitable any time information is presented, because it is impossible to present information without a frame. Still, as the example in this section shows, framing can cause biases. And it is also known from literature (Emby, 1994, Johnson et al, 1994) that the way a problem is stated heavily influences its solution or even its solvability. An example can also be derived from the birthday problem. That could also have been presented as the question: "Give an estimate of the probability that neither of those present at the party have their anniversary on the same date". Then the estimate might have been much better. So there are good reasons to subsume framing under the concept of heuristics, and deal with the corresponding biases.

3.1.4 Accessibility, System 1 or System 2 Judgment, the Affect Heuristic, Prototype or Extensional Attributes

In his Nobel lecture on December 8, 2002 Kahneman reviews the work of Tversky and Kahneman and that of Kahneman and others. By distinguishing two modes of thought, in a "Two System View" for cognition he consolidates the theoretical framework for the heuristics approach. "System 1" is the system of intuition and "system 2" is the system of reasoning. In his lecture he adds a key concept to the theory of heuristics and biases: the concept of "accessibility". He shows that representativeness, availability, and anchoring can be explained by the role the "accessibility" of key features plays in the object to be judged.

In his lecture he gives a more precise definition of a heuristic by calling it attribute substitution and defining it as follows: "A judgment is said to be mediated by a heuristic when the individual assesses a specified target attribute of a judgment object by substituting a related heuristic attribute that comes more readily to mind". This definition does not apply to anchoring effects because here the key mental mechanism consists of temporarily raising accessibility of a particular value of the target attribute, relative to other values of the same attribute.
In his lecture he adds the "Affect Heuristic" to the basic heuristics in judgment. In principle this heuristic explicitly models the influence of affections on judgment. It deals with the extent to which thoughts from system 2 can be corrective for biases for which the judgments from system 1 are susceptible.
He discusses prototype and extensional attributes; prototype attributes being the average of these attributes for a class and extensional attributes being the specific extra of a member of the class for the prototypical attributes of the class. Because of their high accessibility the prototype attributes are the natural candidates for the role of heuristic attributes.

Much of the literature we discuss in this thesis rests on the "classic heuristics" and the framing concept. It adds the possibility of corrections from system 2, for instance by looking at the role of experience and of structuring a task. It seems to us that the idea of accessibility does not give a new opening for research. But the affect heuristic might give interesting new openings: it needs no argument that affections inevitably play a role in assessments.

3.2 Heuristics, biases and validity in risk assessment in auditing

Tversky and Kahneman show that many possible biases lie in wait when people deal with incomplete or uncertain information, at least partly due to the heuristics they use, or that are at work on a subconscious level. If risk assessment by auditors is also subject to these biases, then it can only fail, except for the very improbable event that the biases nullify each other. So the question whether we may expect these biases and if they really occur, is very relevant. Otherwise, if the biases would appear not to be very influential, the question still remains whether risk assessment by auditors is valid in the sense meant in this thesis.
In general, the heuristics that cause the biases are not directly visible when they operate. As stated above, they operate in a human mind and can be deduced from the effects that can be observed, when people deal with incomplete or uncertain information. So when looking for the effects of heuristics and corresponding biases we have to wonder where, in what type of information context, the heuristics will operate. We will do this in the next analysis. It will lead to some expectations, that will be evaluated in our overview of previous research in section 3.4 and be concluded upon in section 3.5.
We have to warn in advance that this analysis, the expectations emerging from it and the evidence we find in literature, do not simply give a justification for our leading research questions. In section 3.6 we will bridge the gap between our findings from literature and these questions.

3.2.1 Biases in auditing due to the representativeness heuristic

Biases due to the representativeness heuristic are very likely to occur. In the first place in a context that is relevant as the normal follow up of risk analysis: forming an audit opinion based on a sample of tests. In general, the size of test samples is small, and

even if other sources of information are taken into consideration when coming to an audit judgment, the weight that will be given to the results of a test sample will be relatively large in general, because with an eye on the truth and fairness of the accounts in general it is deemed to give the strongest information.

Expectation A: Small samples tend to be taken as (too) representative of the population.

On the other hand it can be argued, that auditors make an almost exhaustive use of their prior information on the accounts under audit. This is paramount in the basic assumption of the audit methodology: the audited account is true and fair, the 'only thing' an auditor has to do is to gather sufficient evidence to substantiate that. For instance the Leerboek Accountantscontrole 4/5 (p. 7) states (translated from the Dutch): " The purpose of the audit of the annual accounts (..) is (..) *confirming* the truth and fairness of these accounts with a view of their use in societal intercourse" (see Leerboek, 2003). On the majority of occasions this assumption is true, and therefore it is justified not to reject this assumption on the basis of only little counter information, but only to reject it when there is sufficient counter information.

Expectation B: The representativeness heuristic in auditing does not lead to disregarding prior probabilities of truth and fairness of an account.

Even an over-weighting of prior probabilities can be expected, also due to a less justifiable practice in which the assumption of a true and fair account causes auditors to neutralise information that indicates an unfair or untrue account. For instance by declaring an error found as totally incidental, and therefore leaving it out of the extrapolation. Seen from a purely statistical point of view, this practice violates the assumption of randomness that underlies statistical inference (because the randomness is made conditional on the absence of a certain type of error), but maybe a kind of reasoning can be conceived that neutralises this objection. However, in this light Burgstahler & Jiambalvo (1986) do exactly the opposite: they show why that conditional randomness is fatal for a sound inference and thus raise serious objections against this practice. These objections have lately been convincingly confirmed by Hendriks et. al. (2005). But this practice still lasts, with a danger of an unjustified unqualified opinion.

Expectation C: An over-weighting of prior probabilities can be expected.

The three expectations formulated so far are contradictory to a certain extent. We believe that their effects more or less keep each other in balance, so the net effect will depend on the circumstances; see also the discussion after expectation E.

Bias due to representativeness can also occur due to (lack of) predictability. T&K (1974) define this as the phenomenon that a prediction is made solely on the substantive content of the information, but regardless of the strength of the same information. With valuation of stocks or with an assessment of the continuity of a business it would mean that the same assessment would be given, more or less regardless of the strength and quality of the information on which the assessment is based. For instance, the information could be that the saleability of goods is high; the strength of the arguments that substantiate this information can be of varying quality. It is not very likely that the experienced auditor lets himself be led astray by a lack of predictability.

Expectation D The representativeness heuristic will show no bias due to lack of predictability.

Bias due to misconceptions of regression effects could occur if current values are predicted from past values. An extremely high profit of last year could lead to such a regression effect. It is also plausible that the assessment of IR and CR will show a regression effect on the assessments of the previous year. This is because an auditor

will look at the assessments of last year and wonder how stable the audit context has been during the year and, dependent on the stability, he will closely relate the assessment of this year to that of last year. In particular this mechanism plays a role when, last year, the administrative processes were found to be of insufficient quality; the audit department urges the accounting department to improve these processes, which they do. The audit department observes the improvements, but still as a matter of precaution assesses the control risk at "high". So knowingly the auditor weighs the results of last year too heavy. This boils down to

Expectation E: Regression effects in risk assessment will be found.

Bias due to sample size and bias due to prior information might occur simultaneously. In a case where they have opposite directions, the resulting effect may be a correct assessment. In case they have the same direction, which is the case when the occurrence risk is assessed as low and the small test sample shows no or only tiny errors, the auditor may have a too positive assessment. It needs no argument that this bias is more likely to occur when sample size is smaller as a consequence of a favourable risk assessment.

3.2.2 Biases in auditing due to the availability heuristic

The *availability heuristic* will almost surely be present in the way auditors come to conclusions on the basis of the vast amount of information that they have of their audit object. The first thing that comes to mind is that the *retrievability* of the most recent information will be larger than that of older information. In the literature of audit research therefore the availability heuristic is relevant under the label of "*recency effects*". It also may mean that there is an effect on the conclusions due to the order in which the information is presented or gathered.

Expectation F: Bias due to the availability heuristic will be there in the form of recency effects or order effects.

A more speculative effect may be due to the *imaginability* of possible or plausible combinations of events in the administrative processes that are the object of risk analysis. Two contradictory arguments are relevant with regard to this heuristic. The first is that auditors are educated to be keen on error generating conditions, so it may be expected that a possible bias will not be too large, or even absent. The second is that error generating conditions will evolve or deliberately be changed, due to which imagination falls short and such a condition can be overlooked. Therefore a risk may be assessed too low. Imagination may lead to an endless sequence of possible conditions in which the availability heuristic may lead to biases; we conclude that it is a relevant heuristic for research with respect to fraud. But this means that in regular research on auditing the problem hardly will be addressed.

Bias due to the *effectiveness* of a search set seems unlikely. A search set is the combination of knowledge and information about the audit object, or more general the object of searching, combined with the heuristics used. In general a search set is interpreted as being a mindset: a set of relevant knowledge and heuristics in the person's mind. But in the auditor's case this mindset will be controlled to a very large extent by the methodology of auditing itself. And if, as a consequence, the search set is interpreted as the audit methodology, this leads to questions outside of the scope of heuristics and biases.

3.2.3 Biases in auditing due to the anchoring and adjustment heuristic

The *adjustment and anchoring heuristic* may be very relevant in the aggregation that an auditor has to make, based on all the risk factors he has detected in the administrative processes and their context. It will be hard for him to make a distinction between inherent risk and internal control risk and where to situate certain risk factors. He will have to be aware of the dependency between certain factors. It will make a difference whether he sees the aggregation as conjunctive or disjunctive events, whether he starts with very low probabilities, etc. A problem in dealing with this heuristic will be that the possible risk causes are numerous, and their interrelations and dependencies are often very hard to explore. Moreover, much of the assessment of the risk due to these factors occurs in the head of the auditor, so that it is not even clear which risk factors (including their relationships) the auditor deals with. They can only be deduced, to a certain extent, from the results of his assessment.

In principle the same considerations apply to the decomposition of the possible misstatement and corresponding opinion by looking at the assertion level, albeit that the assertions in general are well-defined (see also the discussion in 3.3.3). Assertions like accuracy, ownership, completeness, etc. are basic to the audit opinion on the truth and fairness of the account. Opinions on the assertion level have to be aggregated in order to come to an opinion on the account level. This aggregation will be subject to the mechanisms and problems discussed above. Lea et al (1992) give a very interesting discussion on the aggregation of risk assessment at the assertion level into an assessment at the account level, which confirms the existence of aggregation problems.

So there is complexity in assessing risks on at least two dimensions. Table 3.1 gives an outline.

Table 3.1: Complexity as assertions by risk factors

assertion → risk factor↓	completeness	ownership	..	assertion m
Complexity	sub-assessment (1,1)	…….	..	sub-assessment (1,m)
professionalism	…….	…….	..	……
…….	…….	…….	..	……
risk factor n	sub-assessment (n,1)	…….	..	sub-assessment (n,m)

The risk factor "complexity" for instance may be representing the expectation that more complex organisations are more susceptible to material errors; the risk factor "professionalism" may be representing the expectation that more professional personnel reduces the susceptibility to material errors. On each of these factors and for each of the assertions the auditor has to assess its influence on the risk of a material error, the occurrence risk. This means that the assessment task as a whole will ask for n times m sub-assessments.

In principle every sub-assessment is of the form: "does this risk factor apply to this assertion and if so, to what extent does it cause risks?". Subsequently these nxm sub-assessments will be aggregated. Anchoring with undue adjustments lies in wait.
So the relevance of this heuristic is without question, but the way it operates is very hard to predict.

> *Expectation G: The anchoring heuristic will show effects but its direction is dependent on the specific situation; in particular it will show effects when disjunctive or conjunctive events are involved,.*

3.2.4 Biases in auditing due to framing.

In our introduction of the concept of framing we already observed that framing is inevitable anytime you present information. So in the context of auditing the question is not whether framing will have influence on the judgment and assessment of the auditor, but how. Effects of framing already can be observed when the basic task of auditing is considered. For many auditors their mission is to confirm that the annual accounts as issued by management are true and fair, because otherwise they would not have issued these statements (see: Leerboek 2003, p.7). For other auditors their mission is to come to a conclusion regarding the question whether the possible errors found in the annual accounts are not so large that they prohibit an unqualified opinion. In principle these two views are equivalent, but in practice they may lead to different actions and therefore to different outcomes. Many other examples of possible framing effects could be given, they all lead to

Expectation H: Framing effects will be found in auditing.

3.2.5 Conclusions as to heuristics and biases in risk assessment

The previous discussion of the relevance of heuristics and biases as stated by Tversky & Kahneman inevitably leads to the conclusion that they are very relevant for risk assessment by auditors, and even for the way they come to an audit opinion. Of course this has been recognized so far by many researchers, and results of their investigations will be discussed in section 3.4. The expectations as stated above will serve as a guide in interpreting the literature, without claiming completeness of all the relevant literature for a specific expectation. This means that neither denial nor confirmation of an expectation should be seen as a "proof", but at its best as a certain tendency.

At the same time an interesting question will remain unanswered, namely: “Is the absence of biases as discussed in this chapter sufficient for a valid assessment of risks, as meant in chapter 2?” When trying to answer this question, it must be observed that the first step in risk assessment will be the assessment of the risk associated with primitive events or conditions in the administrative processes. It is very unlikely that each of these primitive events can be assessed at the "right" risk. For instance, it will be hard to assess the risk associated with the quality of a set of application controls or of a set of user controls. In the first place it is very hard to separate them from the other controls and the environmental controls and even if this is possible the task remains to assess the associated risk. And only then the aggregation of these risks to an overall risk of the existence of a material error in the accounts can take place, with the accompanying possibilities of biases.

The conclusion can only be that validation on the criteria of chapter 2:

- the ‘audit position’ of the error rate,
- the error rate,
- the sampling risk and
- the distribution of the error rates conditional on the risk assessment,

remains necessary. Literature on research with this orientation will also be discussed in the current chapter.

3.3 The validity of the audit risk model

So far we have discussed the validity of risk assessment from a general point of view and with a specific view on the possibility that biases in risk assessment might occur. We have seen biases conceived in literature and their existence confirmed (bias will

specifically and more thoroughly be examined in section 3.4 for audit situations), but the question is still unanswered whether this implies a lack of validity with respect to our validation criteria. This validity still might be found. But even if our study would show this validity, invalidity of the ARM as a model could prevent its application to be valid. It could, for instance, be that OR is assessed via IR and ICR, without taking into consideration their dependency. Therefore it is relevant to look at the quality of the model itself. We will do this from three perspectives:

1 the event structure of the model;
2 the statistical validity of the model;
3 the level at which risk is assessed.

3.3.1 The event structure of the ARM.

Four events are distinguished in the ARM (see Panel 2000 p.164; SAS No. 47):

- the coming into being and existence of a material error if the related controls in the business processes would not work; the inherent risk (IR) is the risk of this event happening;
- the failure of the related controls in the administrative processes to prevent or detect and correct a material error on a timely basis; the control risk (CR) is the risk of this event happening;
- the failure of analytical review by the auditor to detect a material error; the risk of analytical review (RAR) is the risk of this event happening;
- the failure to detect a material error in a sample of tests of detail; the sampling risk (SR) is the risk of this event happening.

The precise definitions of these risks are given in section 1.4.
Table 3.2 shows the context of business and administrative processes, in particular its controls and the actions the auditor can take in the form of analytical review and substantive tests of detail. It also shows the resulting events and the related risks, as far as relevant in the ARM. It is clear that the four resulting events: that of the possible existence of a material error, are distinguished because of their distinct places in the business, accounting, and auditing processes.

Table 3.2: Actions, events and the risk in the ARM.

Action of audit object	Action of auditor	Event	Action of auditor	Assessed risk
Business processes		Creation of material error?	Assessment of quality of processes and environment	Inherent risk, IR
Set of controls		Not preventing or detecting and correcting material error	Assessment of quality of controls	Control risk, CR
Annual accounts	Analytical review	Not detecting a material error	Assessment of quality of analytical review	Risk of analytical review, RAR
Annual accounts	Tests of detail	Not detecting a material error	Assessment of quality of tests of detail	Sampling risk, SR
			Combining the four risks	Audit risk, AR

The problem with the events as defined above, is that the causes of the first and the second instance of the possible existence of a material error and therefore IR and CR, are very hard to separate. This is caused by the fact that many controls, such as separation of duties, authorisations, competence level of employees, etc. are also part of, or help determine, the business processes, which cannot be imagined to work without these controls. So it is almost impossible to assess inherent risk apart from the control risk. Waller (1993) points at the preventive qualities of the internal controls. They precede the coming into being of errors and therefore directly influence the IR, but they cannot be modelled in the ARM. And where to model the competence level of employees: is that preventive (IR), just part of the context in which the business processes operate (IR), or detective (CR). Ambiguities in this respect are likely to cause inconsistent assessments of the IR and CR (see again Waller 1993).

In this thesis we circumvent this problem of separability by validating the occurrence risk, the combination of inherent and control risk. This is also justified with a view on the statistical validity of the ARM (as a result of the lack of separability), as we will explain in the next subsection.

3.3.2 The statistical validity of the ARM

In the ARM the four composing risks are multiplied, which implies that they are assumed to be independent in the statistical sense. This means that irrespective of the assessed IR, the CR is assessed. But the fuzzy event structure as explained above already causes dependencies between IR and CR, as does a possible anchoring effect: if the inherent risk IR is assessed as "high" this may cause an upward tendency in the assessment of the control risk CR (Waller 1993). But also the opposite effect may happen: an as "high" assessed IR, may cause the auditor to start from the idea that the internal control will be strong and subsequently assess the CR as "low" (idem).
The dependencies are circumvented when IR and CR are aggregated into a new risk: the occurrence risk (OR, again Waller 1993, HCDAD 1997. Kinney, 1992) or when the ARM is changed into: AR = f (IR, CR, RAR, SR).
It is obvious that OR = f (IR, CR) is included in the previous expression.

In the Netherlands the statistical validity of the ARM was extensively discussed in the end of the 80's and the beginning of the 90's of last century (see, among others: Ten Wolde (1989), Veenstra & Van Batenburg (1990), Schilder (1991), Broeze et. al. (1991), Lammerts van Bueren (1991), Hoogewoning (1991)).

It is obvious that as long as the statistical validity is questionable, valid assessment of OR, or of IR and CR, are only a necessary and not a sufficient condition for a valid application of the Audit Risk Model. Here "valid" has a broader meaning than it has in the research in this thesis. A valid model refers to an application of the ARM that leads to the correct size of substantive testing and other substantive procedures, given the assessed occurrence risk. The size is correct when the reliability of the audit opinion has the "nominal level". "Nominal level" is the level of reliability that was aimed at by the audit design; it is the level of reliability, or complementary the risk, at which the size of substantive testing is calculated. In a valid model, this nominal level will be equal to the actual reliability.

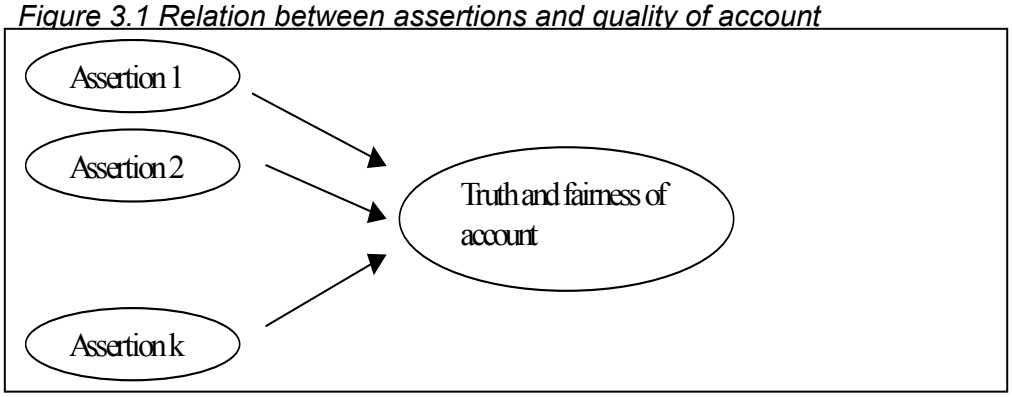
The validity investigated in this thesis, only refers to valid assessment of the occurrence risk. In principle this does not suffer from the statistical invalidity of the audit risk model, unless the occurrence risk were calculated from the assessments of IR and CR by multiplying them. Then this would be incorporated in our data on OR, because we asked our respondents for their OR assessment. But in the discussion we had with our

respondents they all reported their assessment of OR to be based on a judgmental combination of IR and CR.

3.3.3 The level at which risk is assessed.

In general two levels are in use at which some component of audit risk can be assessed: (1) the assertion level, (2) the account level. At the assertion level "assertions" like completeness (have all transactions been recorded) or existence (does an asset exist) or accuracy (is the recorded amount correct) are the object of risk assessment: the risk is assessed that some material error might occur in this assertion. An assertion can also be seen as an aspect of the truth and fairness of the account. So the assertion level does not imply a subdivision of the account in smaller sub-accounts, but a subdivision of the quality of the account in aspects. Assertions are also labelled as "audit objectives": an audit activity can have the objective of assessing the completeness of a sub-account, or of the existence of a debtor, etc. Related to such an audit objective or assertion is an opinion on that assertion. It is clear that the risk at the account level is some aggregate of the risks at the level of the assertions. Figure 1 sketches the relevant relations.

Figure 3.1 Relation between assertions and quality of account



When the auditor can give a positive opinion on the assertions and if the aggregate of the errors in the assertions does not exceed the level of materiality for the account, the auditor can give an unqualified opinion of the account.

The aggregation of the evidence found with respect to the assertions, both with respect to the size of the error, and with respect to the associated risk, should be done in a very systematic way (see Srinindi & Vasarhelyi 1986, Lea et al, 1992). In practice no formal calculus to execute this aggregation is used and there are only poor heuristics in use to achieve such an aggregation, like: “take the risk associated with the most important assertion”. (See again Srinindi & Vasarhelyi 1986 and HCDAD 1997). In 3.2.3 we discussed that risk assessment at the account level will be composed in principle of risk assessments at the level of the assertions, which means that the absence of valid aggregation rules is a serious problem for the assessment at the account level. At best an auditor will resort to letting himself be led by heuristics dealt with earlier in this chapter, but this will mean that he will also be subject to the corresponding biases. This will also mean that it is highly probable that risk assessment at the assertion level will show greater validity than risk assessment at the account level (see Waller 1993).

In the above discussion, one simplification has been allowed which might complicate the relation between assessments at the assertion level and assessments at the account level. In the discussion we took the level of materiality at the assertion level for granted. Maybe this implies that, for every assertion, the same level of materiality will be adopted, say, equal to the level of materiality for the whole account. But this is far

from self-evident: when for instance, the truth and fairness of the account depends on three assertions, and errors on these assertions are additive when aggregation to the level of the account takes place, in principle one-third of the materiality level for the account could be material for the assertion level. Many variations on this complication can be imagined and they might cause the expectation to fail that risk assessment at the assertion level may be more valid than risk assessment at the account level as expected by Waller (1993).

When we realise that risk assessment at the account level is almost inevitably aggregated from assessments at the assertion level (implicitly or explicitly) it is clear that there are more problems with the ARM than those of heuristics and biases in judgment under uncertainty.

3.3.4 Conclusions as to the validity of the ARM

This very short account on the validity of the ARM serves to show that a valid risk assessment is necessary, but not sufficient for a correct application of the ARM. Especially when IR and CR are assessed separately, a valid assessment still can lead to incorrect results due to weaknesses in the ARM. In this study this is taken as a limitation on the subject. Our point of departure is that the first thing to be sure of is a correct assessment of the risks and only when that is accomplished, does improvement of the audit risk model make sense. This means that we choose a course of action that is different from, for instance, Mollema (2003, 2004), who aims at improving the model without making a problem of the validity of risk assessment. Our looking at the validity of IR and CR at the account level also means that we do not consider problems of aggregation, neither from assessments on risk factors nor from assessments on assertions. Only in chapter 7, where we deal with risk indicators, we will meet the problem of aggregation, but there it is not our object of study.

3.4 Relevant literature

We have chosen the work of Tversky and Kahneman (T&K) as a framework for our study of the literature on risk assessment. T&K give a very interesting frame for understanding what happens in risk assessment. As a consequence many studies on risk assessment can be described within this framework. Moreover many studies explicitly refer to the T&K concepts. So for an overview of research into risk assessment the T&K paradigm offers opportunities to come to a well-fitting categorisation.

At the same time not all relevant literature can be subsumed under this framework. So in this subsection we will extend our search beyond the T & K-framework. Before we give an account of the relevant literature, we notice two contingencies with our research problem.

> Firstly we observe that complexity of information may cause the same heuristics to operate as discussed in sections 3.1 and 3.2. This is because complexity can be a source of uncertainty or incompleteness in or of the information. Information can be so complex that relevant interrelations, interactions and effects are beyond the information processing capacity of the human mind and/or even beyond the possibilities of more formal modelling. Complexity possibly causes effects of information to be subject to order effects and consequently availability effects, anchoring, etc. This observation results in the conclusion that literature on the processing of complex information will also be relevant for this thesis.

Secondly we recall our discussion in section 2.4: the assumption that system tests (tests of control) have a predictive value for the error rate or the rate of misstatements. In our study we will check this assumption, and therefore literature on the relation between tests of control and/or compliance testing on the one hand and the error rate on the other is relevant too.

We will order our overview of literature as follows:

3.4.1 Studies that are directly related to the heuristics and biases paradigm
3.4.2 Studies that investigate consistency of risk assessment with some criterion
3.4.3 Studies that deal in some sense or another with the complexity of the object of risk assessment
3.4.4 Studies that investigate the value of system tests for validation of risk assessment.

3.4.1 Studies on the heuristics and biases paradigm

3.4.1.1 Framing

Johnson et. al. (1991) and Emby (1994) investigate the effect of 'framing' on the assessments of an auditor. As stated in section 3.1.3 *'framing'* refers to the fact that information is always given within a point of view or frame of reference. This frame is somehow given to, or adopted by, the one who performs the assessment. In audit research the dependent variable may be the quality of some business, or the audit effort to be given by the auditor. Bias towards the frame of reference as an effect of framing is shown in both studies, but experience or extra information may counteract this bias.

Johnson et al (1991) state that management creates a possibly misleading frame by issuing financial statements that suggest a growth firm, when in fact a fraud is hidden in the statements. Auditors, experts and novices were provided with extra information about this firm in which a fraud case is hidden. Indeed it appeared that the idea of growth firm serves as a frame for further investigation. But Johnson et al found that subjects with sufficient knowledge about the industry were able to change the growth frame into a fraud frame and consequently detect the fraud. This was independent of experience as an auditor as such.

In a second experiment subjects were provided with a case with a financial misstatement. Expert auditors appeared to be able to look through this frame induced by the financial misstatement and detect it. Novices did not succeed in that task. In this second experiment knowledge of the industry did not seem to have influence on the success with respect to the task

The effect of framing was shown; it was also shown that expertise, be it as an auditor or as an expert in the industry can cope with this framing effect.

A brief comment is also due: the phenomenon that Johnson et al investigate and label as "framing" may also be seen as "anchoring". The experimental subjects show a tendency to cling to a given starting picture.

Emby (1994) presented the same assessment problem in two frames. In the first, subjects were asked *to assess the risks* in an internal control system given in a brief description, in the second, subjects were asked *to find the strengths* in the same internal control system with the same brief description. In a third, control, group the same information was given but without framing questions. All groups were asked to decide on the level of substantive testing that they would plan in the given conditions. Emby found no significant differences in the planned levels between the three groups. The three groups were provided with more information in a sequential way or in a simultaneous way, giving a three by two design with the frame factor. Now Emby found a significant difference in chosen level of substantive testing between the risks group

provided with sequential information and the strength group, provided with the simultaneous information. It means he found an interaction between framing and recency effects.

> *Intermediate conclusions as to framing*
> *Framing appears to have influence on the auditors assessment. Expectation H is confirmed.*

3.4.1.2 Anchoring and adjustment
O' Donnell & Schultz (2005), Wilks (2002), Butler (1986), Joyce & Biddle (1981a), Krug Nelson (1995), Smith & Kida (1991) and Tan (1995), (among others) investigated the effect of 'anchoring and adjustment', or of phenomena that we deem to be very similar to anchoring. As stated in 3.1.2.3, '*anchoring'* is the heuristic which reduces complexity by taking a starting picture ('anchor') as the picture to be '*adjusted*' to comply with (new) information. The dependent variable is the assessed risk or probability.

O'Donnell & Schultz (2005) investigate the 'halo-effect', for the first time described by Thorndike (1920). The halo-effect is the mechanism that a favourable (or unfavourable) judgment of an object on some attribute is likely to lead again to a favourable (or unfavourable) judgment on some other attribute. O'D & S find this halo-effect: when auditors assess the account-level risk: they tend to adjust this to the strategic risk as it is assessed in the business process analysis. The strategic risk apparently serves as an anchor for the account level risk.

Wilks (2002) investigates the effect on the tendency to agree with supervisors' views of a subordinate who is aware of these views. This would be caused by distorting evidence or by evaluating evidence in a way consistent with the supervisors views. In an experiment Wilks finds that auditors who learn the partners' views before evaluating the evidence, evaluate evidence and make going-concern judgments more consistently with the partners' views, compared to auditors who learn the partners views after the evaluation phase.
A brief comment is at its place: Wilks' findings can be seen as a matter of anchoring of a judgment; obviously compliance to desirable behaviour may also give an explanation of his findings.

Butler (1986) gave the result of an audit sample to his experimental subjects and asked them to assess the audit risk (AR), given this sample result. In the presentation of the results of the audit sample Butler manipulated the information on the allowed audit risk. Clearly this allowed AR is irrelevant for the actual AR, but Butler expected this extra information to work as an anchor. His experiment confirms this expectation.

Joyce & Biddle (1981a) introduced an anchor in their experiment by manipulating the starting question. In the first experimental condition they asked their experimental subjects: what is your estimate of the number of fraudulent businesses, is it more than 10 per 1000 and how much more? In the second experimental condition they asked their experimental subjects whether their estimate was more than 200 per 1000 and how much more or maybe less? This manipulation caused significant differences in the given answers by the two experimental groups.
In a second experiment they found that the sequence in which information with respect to the internal control was given also led to anchoring effects: experimental subjects provided with prior information, followed by negative adjustments chose a higher level of audit effort than experimental subjects provided with the same information (prior and adjustments) at once. Auditors with more experience (experience varies between 0 and 27 years) showed less bias by the anchoring effect.

Again a brief comment is worthy of note: the phenomenon that Joyce & Biddle investigate in their experiment and label as “anchoring” may also be seen as “recency”. Seen as “anchoring”, the experimental subjects over-adjust in the sequential presentation, which is inconsistent with the anchoring and adjustment ideas. The recency heuristic directly explains the experimental effect.

Joyce & Biddle (1981a) also investigated possible differences in assessments of probabilities which result from a combination of primitive probabilities in the case of conjunctive and disjunctive events. In their experiment the primitive probabilities were known to the experimental subjects; the same assessment problem was presented in a conjunctive (condition 1) and a disjunctive (condition 2) mode. Joyce & Biddle did not find differences between the two conditions in the assessed risk.

Smith & Kida (1991) showed anchoring effects when risks are combined in a conjunctive or a disjunctive presentation of a problem. They found that risks are overestimated with a conjunctive presentation and underestimated with a disjunctive presentation of the combined event. They explained this effect from the assumption that the probabilities associated with the primitive events serve as an anchor when the probability of the combined event is estimated. (Remember the “birthday problem”: it was presented in 3.1.2.3 as a disjunctive event, the relevant probability is underestimated).

Krug Nelson (1995) found that adjustment into a posterior probability of a given prior probability of a material misstatement, based on the given results of an audit sample, may be insufficient. The prior probability appeared to serve as an anchor.

Tan (1995) investigated the effects of repeat engagements: is there more “consistency” in the assessment in repeat engagements compared to new engagements? By “consistency” is meant a lack of variability in the aspects that are deemed to be relevant for risk assessment. Tan found such consistency effects, which also can be called anchoring effects. He also investigated the possibilities of mitigating that consistency effect by staff rotation or by inducing review awareness. He showed that this awareness, raised by an expected review, leads to more vigilance and therefore attention for inconsistent findings. Also, job rotation leads to more attention for inconsistent findings. In addition, a panel of audit partners found the decision process under the rotation and review awareness conditions to be more appropriate.

> *Intermediate conclusions as to anchoring*
> *There is a strong tendency for an anchor to prevail in the final assessment; expectation G is confirmed.*

3.4.1.3 Representativeness

Bar-Hillel (1979), Joyce & Biddle (1981b), Nelson (1995) and Smith & Kida (1991) (among others) explored the effect of ‘*representativeness*’. As stated in 3.1.2.1, ‘representativeness’ is the heuristic which reduces complexity by taking given information as representative for the population, irrespective of relevant base rates, sample size, or quality of the information. The bias due to this heuristic was found by Nelson (1995), Joyce & Biddle (1981b) and Bar-Hillel, (1979) both where it regards not taking into account the sample size, and where it regards neglecting a base rate in which a phenomenon occurs. Smith & Kida (1991) found that the bias is smaller with more experienced auditors and with more realistic cases, administered to the experimental subjects.

The studies of Burgstahler et al (2000) and Messier et al (2001) are not explicitly related to a heuristic as discussed; but their findings clearly relate to the

representativeness heuristic. They show that people let themselves lead by this implicit principle which causes bias in their assessment.

Messier et al (2001) found that the use of a recent AICPA 'Audit Sampling guide' leads to sample sizes well below what would be necessary had a statistical sampling approach been used. Their study might even mean that use of this Sampling Guide guides in the direction of under-auditing. They are concerned that greater assurance is inferred than is jusitified by the sample (had it been statistical). We add to their conclusion that the representativeness heuristic gives a perfect explanation for the effect they found; see also expectation A (3.2.1).

Burgstahler et al (2000) found that auditors tend to underestimate the effect of both projected error and uncertainty when evaluating the aggregate error and the need for adjustments to financial statements. They even misjudge the meaning of an upper error bound (statistically derived) exceeding the materiality. So there is a basic inconsistency with the laws of probability. Especially their results with respect to the upper error bound, can be explained by the representativeness heuristic.
Burgstahler et al present a critical review of the existing research literature of expertise in auditing. The review is organised around two approaches: the behavioural and the cognitive approach. Results from studies using the behavioural approach indicate that expert auditors do not behave differently from novice auditors. Results from studies using the cognitive approach are more encouraging. They indicate that there may be knowledge differences between experts and novice auditors and that these differences might lead expert auditors to use decision processes that differ from those used by novice auditors.

> *Intermediate conclusions as to representativeness*
> *The representativeness heuristic shows clear effects, both as to disregard of sample size (see expectation A; confirmed) and as to not taking into account the base rate (see expectation B, not confirmed and C, confirmed). Expectation D & E were not dealt with.*

3.4.1.4 Availability
In 3.1.2.2 we stated that the *availability heuristic* makes people assess the frequency of a class or the probability of an event dependent on the ease with which instances of occurrences come to mind. It is investigated by (among others) Ashton & Kennedy (2002), Hall et al (2001) and Cushing & Ahlawat (1996).

Ashton & Kennedy (2002) show recency effects in going concern judgments. When this judgment is based on a so-called 'step by step procedure' (form a judgment after the first information item and revise this after each new information item), a bias towards the most recent information is shown, compared to a judgment given only once, when all information has been processed, the 'end of sequence procedure'. Moreover, A&K show that proper self-review eliminates the recency effects.

Hall et al (2001): investigated whether doubling the sample size mitigates the selection bias in haphazard sampling that is (or may be) caused by 'call properties' of units, as size or isolation (the larger and/or more isolated are more likely to be selected). So they studied whether bias caused by the *availability heuristic* can be mitigated by increasing, doubling the sample size. They found that the bias clearly exists (some 30% over-representation), and also that the doubling hardly has any effect on the bias.

Cushing & Ahlawat (1996) found that recency effects, a form of availability effects (see 3 2.2), disappear when the auditor is asked to document the evidence he bases his opinion on.

Intermediate conclusions as to availability
The availability heuristic shows clear effects; which in some cases can be circumvented by proper documentation or self review (expectation F only partly confirmed)

Intermediate conclusion of 3.4.1: heuristics and biases

An intermediate conclusion of this sub-section can be drawn as: the heuristics conceived and investigated by Tversky and Kahneman (1971, 1974, 1981) are fruitful paradigms for investigating many aspects of risk assessment by auditors. In general, ample evidence is found for their existence, especially because the corresponding biases are found in many experiments. An encouraging finding for the professionalism of auditors is that on several occasions experience as an auditor appears to help to prevent biases. In various experiments probabilities or risks of a quantitative nature were used; in these cases formally speaking risk assessment also was validated. But the question whether an auditor is able to derive valid risks from his assessments of real-life administrative processes, has not been addressed.

3.4.2 Studies on consistency of risk assessment with some criterion

We organise the studies in 4 sub-subsections: consistency (1) with audit design, (2) with other subjects (consensus), (3) with error rate or another indicator of the assessed risk, and (4) with size of errors.

3.4.2.1 Consistency with audit design.
Basu & Wright (1997), Dusenbury et. al.(1996), Mock & Wright (1995), Joyce, E. (1976), Gaumnitz et al (1982), Srinindi & Vasarhelyi (1986), Houston et al. (1999), and Elder & Allen (2003) (among others) deal with consistency of risk assessment with audit design.

Basu & Wright (1997) performed an experiment in which eight control-environment factors ((1) management philosophy and operating style; (2) organisational structure; (3) audit committee; (4) methods of assigning authority and responsibility; (5) management control methods; (6) internal auditing; (7) personnel policies and procedures; (8) external influences see SAS 55), contained in the US professional auditing standards, were manipulated as either positive or negative for three different client sizes. Also the consistency of an auditor's control risk assessment with the formulation of the preliminary audit strategy was investigated. It was shown that auditors do not place equal emphasis on the eight factors. However the reliance placed on these factors is not significantly different across clients of different size. A significant impact of risk assessment on the planned audit strategy was found.
An interesting feature of this study is that it is one of the rare studies on separate risk factors. We saw already that the complexity of the assessment task of an auditor is caused among other things by the necessity to aggregate risk assessments on separate assertions into one risk assessment on the account level. This article deals with the complexity on another dimension: it studies factors in which the causes of risk itself can be decomposed. The introduction of the risk factors by Basu & Wright expands the complexity that is already caused by the distinction between assertion and account level. We discussed this complexity in 3.2.3 in figure 1.

Dusenbury et. al.(1996) investigated the consistency, by way of an experiment, of three ways of assessing IR, CR and APR (analytical procedures risk) to come at an allowed TDR (test of details risk), namely by way of the (classical) ARM, by way of a firms model and by way of a belief based model. The latter uses Dempsters rule to combine beliefs and bases the degree of belief on the quality of the information. It appears that

the belief based assessment is the most conservative, then that of the audit firms and then the ARM. It is claimed that this order is dependent on the type of material. This means that the order could be the other way round, with some other type of material.

Interesting quote
The consensus of the extant research is that this specification of the (SAS) model is deficient. First the target audit risk selected ex ante is generally less than the actual audit risk achieved ex post, when the target risk is used to determine the allowed tests of details risk (..) audit plans based on this model may not be as conservative as they should be."
end of quote

It will appear in our research that the same lack of success in keeping audit risk under the desired level, occurs rather frequently (we will call this 'ineffective audit'). This study indicates that there is no consistency over assessment methods and no consistency of risk assessment with audit plans.

Mock & Wright (1995) replicated their extensive previous archival study of evidential planning by audit teams (Mock & Wright, 1993). Data were gathered on risk assessments and evidential plans in the accounts receivable area from the working papers of 76 randomly selected clients in two industries.
Corroborating prior archival studies, the results did not indicate a strong statistical association between changes in assessed risks and audit plans. Programmes were found to change little over time with many tests done across a broad array of engagements. In addition, risk assessments were not found to differ significantly between the two industries. Moreover the planned extent of testing was positively related with the number of prior errors and there was some evidence that programme plans were related to risks at the assertion level. In general there was a lack of sensitivity of audit programmes to risks.

Joyce., E. (1976) notes that "One of the difficulties involved in studying the validity of auditors' judgments, is the absence of a suitable criterion by which to distinguish correct from incorrect judgments". The solution we found for this problem in chapter 2 implicitly confirms Joyce's remark: we only can validate judgments in a statistical way, by looking at the quality of a set of judgments. In principle, chance can never be studied in single cases. At best the error rate, or rather the sampling risk can be seen as a single case indicator. But in this thesis, these are also used in analyses of correlations, so for multiple cases.
As a result of his review of the literature on this subject, Joyce states that individual auditors' judgments usually show a high variation across auditors. The only exception Joyce found is in Ashton (1974); he asked his subjects to rate the strength of the internal controls. Subjects in other studies made judgments about the amount of audit work to perform. To the extent that different firms have different internal procedures, continuing education programmes, etc., agreement among auditors within firms would be expected to exceed agreement among auditors between firms. This expectation was confirmed with the exception of one firm.

Gaumnitz et al (1982) found that auditors achieve consensus in their assessment of the internal control strength and the audit program planning tasks. An explicit, quantified assessment of the internal control was asked. This may be the difference with Joyce's (1976) study where internal control was not assessed that explicitly. In Joyce's studies planned audit work expresses the assessment. Possibly as a consequence Gaumnitz et al also find consistency in the inverse relationship between the internal control assessment and the audit program planning tasks.

Srinindi & Vasarhelyi (1986) attempt to reconcile the apparently discrepant research findings of Gaumnitz et al (1982) and Joyce (1976) by dividing the process of risk assessment into three different stages:
1)identification, 2) evaluation, and 3) interpretation.
The third stage, where auditors decide on substantive tests based on their perception of the internal control strength, was analysed in descriptive terms. A high degree of consensus between auditors was found in substantive test planning decisions when numbers were provided representing the system reliability. At the same time, large divergence was observed when only *component* reliability values were provided in the form of a number. Therefore, the experimental results indicate that auditors will disagree on how to aggregate audit evidence but, once one aggregation rule is established, high consensus will follow. With these findings Srinindi & Vasarhelyi reconcile the above-mentioned, seemingly discrepant findings: variance in risk assessments might very well be due to varying aggregation rules. Their findings therefore also lend credibility to the need and desirability of using internal control reliability *decision aids*, that provide a consistent algorithm to aggregate the component reliabilities.
It can be concluded that the consistency question is more complex than at first sight. Assessed risk and substantive test planning decisions are shown to be consistent when an assessment for the system is given in numbers, without need for aggregation from components. But when aggregation of risks (in numbers!) from components is needed, this is done in a very diverging way by different auditors and this leads to inconsistent substantive test planning decisions over auditors. So the study offers another example of the problems with risk assessment when it regards larger or complex entities, where implicit or explicit aggregation of risks or probabilities has to take place.
Moreover, in practice, the assessment of the quality of internal control still has to be transformed into a quantitative assessment. In this transformation there also is a variability between auditors and cases, which, as S&V show, may very well be enlarged in the combination of assessments on components.

Houston et al. (1999) identify conditions under which the audit risk model does, and does not, describe audit-planning (investment and pricing) decisions. In an experiment, audit partners and managers examined one of two cases where a material misstatement was discovered. In one case the misstatement was a possibly material error, in the other case the misstatement regarded an overt irregularity: the inventory system was inconsistent with GAAP. The auditor assessed the elements of the audit risk model and assessed the business risk. Based on his assessments he provided recommendations for the audit investment and fee. When the likelihood of an error was high (according to the assessment of the experimental subject), the audit risk model performed better than the expected business risk (see also 2.2.1) in the explanation of the audit investment. Moreover the client was not charged a (business-) risk premium. When the likelihood of an irregularity was high (again, according to the assessment of the experimental subject), the expected business risk performed better than the audit risk model in the explanation of the audit investment, and the fee the (hypothetical) client was charged, contained a (business-) risk premium. These results (n=17 in each condition) suggest that the ability of the audit risk model to describe auditor behaviour and the inclination of auditors to charge a risk premium depend upon the nature of the risks present in the audit. In the presence of errors, the audit risk model adequately described audit planning decisions; in the presence of irregularities it did not.

Elder & Allen (2003) examine changes in risk assessment and sample size decisions between 1994 and 1999. They found a tendency for greater reliance on controls in the later period and lower assessments of IR. Firms that used larger sample sizes in the first period show a tendency for smaller sample sizes in the later period. They find a significant relationship between the inherent risk assessment and the sample size

decisions; this relation was stronger in the first period and not significant for all firms. The relation they found between control risk assessment and the sample size is much weaker.

> *Intermediate conclusions as to consistency with audit planning*
> *Once there is a result in the form of an occurrence risk (or inherent risk or control risk), the decisions with respect to audit planning are consistent with this assessment, possibly conditional on the type of error the auditor expects. But as soon as some aggregation of partial assessments, or an assessment of risk associated with the observed quality of an administrative process is involved, consistency with audit planning tends to fail.*

3.4.2.2 Consistency with other auditors (consensus)
Another type of consistency is agreement between auditors on risk assessment. Stone & Dilla (1994), Amer et al (1994), Trotman & Wood (1991), Reimers et al. (1993) and Davis et al. (2000) (among others) study this type of consistency/consensus. As in the previous section, in some studies not only the consensus is investigated but also its consequences for audit design.

Stone & Dilla (1994) investigated the influence of experience of auditors on their risk assessment, next to the effects of the representation of this assessment. They expected experienced auditors to have developed more complex and complete classifications of "domain stimuli" (signs and signals from the audit object, possibly indicating a relevant feature for the audit opinion) than have novices. They state that a response classification in numbers allows for more precision than a linguistic classification. Therefore they expected experienced auditors to show a gain in consensus in a classification of risk when this classification is in numbers rather than in words. This gain in consensus was expected to be absent when the same judgment task was given to inexperienced auditors. In addition, they expected more variation in risk judgment with experienced auditors, due to the unique professional experiences that they have. In contrast, they expected judgment of inexperienced auditors to have more dependency on their education and therefore show less variability; or, in other words, more consensus. These subjects lack the specialised knowledge to improve their judgments when using numbers. In two experiments their expectations were confirmed.
This study is inconsistent with Srinindi & Vasarhelyi (1986), who found: consensus in risk assessment or classification is weak or absent, when there is too much complexity in the task (see end 3.4.2.1).
The two studies to be discussed below find possibly conflicting results.

Reimers et al. (1993) noticed that the Statements on Auditing Standards (SAS) nos. 39, 47, 55 suggest that the assessed control risk has a direct effect on the amount of substantive testing required in the audit. They found that numerical risk assessments differ from linguistic assessments in two ways. First, the assessments of those responding numerically were significantly lower than of those responding with linguistic categories. Second, there was a consistently higher level of agreement among those who responded in linguistic categories. Unfortunately Reimers et al. did not include experience as a variable; therefore it is not clear whether these findings really conflict with those of Stone & Dilla (1994).
Quote: 'As in most audit judgements there is no way to determine correct control risk assessments.' End quote
We hope to disprove the quoted statement to a certain extent in this thesis. At the same time it gives a justification for our way to validate OR in our research.

The study of Amer et al (1994) adds to this possible lack of consistency: it shows that the interpretation of probability phrases differs substantially between auditors and also that auditors are not aware of these differences. So the same object of risk assessment leads to substantially differing assessments by auditors. Moreover it shows that most of the risk assessments are not valid, which raises the suspicion that the few valid assessments in their experiments are only a matter of chance.

Trotman & Wood (1991) conducted a statistical meta-analysis on 17 studies into the consensus regarding the assessment of internal control risk and regarding the decision as to the size of the substantive audit effort. With both variables they found a reasonable (approx. .60) and significant ($p<.05$) correlation. So, according to this meta-analysis, decisions of auditors as to the use of substantive audit resources are consistent with their assessment of risk.

Davis et al. (2000) examined the relation between consensus (estimates coincide to a high degree) and accuracy (estimates are close to the quantity to be assessed), using an error frequency estimation task for which the auditor's overall accuracy is known to be low to moderate. They used real life cases in which the error rates were known. They also examined the extent to which experience had influence on the strength of the relation between consensus and accuracy for the three industries they examined: manufacturing, natural resources and banking. Accuracy appeared to be positively related to consensus for all auditors in manufacturing and for auditors with more than 12 (36) months of experience in natural resources (banking). For banking and natural resources, evidence was provided that auditors with little experience in these industries use a heuristic consistent with manufacturing error frequencies as an 'educated guess' for the specialized industries' error frequencies. This heuristic leads to consensus among auditors, but results in low accuracy. More research is needed.

> *Intermediate conclusions as to consensus*
> *Consensus on risk assessment between auditors on the same object appears not to be stable; it is dependent on experience and representation of the assessment object. Consensus is no guarantee for accuracy.*

3.4.2.3 Consistency with error

In this thesis the most interesting consistency is that of risk assessment with the error rate or another indicator of the assessed risk. Only a few studies deal with this consistency in one way or another. Many studies regard the occurrence (also frequency) of error and not its size.

To a certain extent the study of Davis et al (2000), discussed at the end of 3.4.2.2, can be seen as investigating the auditors' ability to assess risk, although it is not directly aimed at the relation between risk assessment and error. Still the findings of Davis et al are very interesting for risk assessment, because they show that auditors can give reasonable estimates of the occurrence of errors to be expected. And it almost speaks for itself that this is directly relevant for risk assessment.

Roberts & Wedemeyer (1988) did not directly investigate the characteristics of risk assessment. But they did investigate relations that are very similar to the question of this thesis.
In their introduction they note that *"Accurate prediction of the distributional characteristics of errors in financial statements is of critical concern to the practicing auditor. These characteristics directly determine the auditors allocation of effort and, eventually, evaluation of the implications of the results of that effort"*. They describe the role of risk assessment, without needing that concept, because they translate it into the

distributional properties of the error in the audit object (see the parallel with our discussion in 2.2.4 and 2.3.5).
They found six general attributes that can be used to predict whether an audit engagement is likely to have a significant amount of monetary error:
1. the control environment;
2. employee integrity;
3. financial strength, particularly liquidity;
4. complexity, particularly unusual transactions;
5. regulation of accounting and reporting practices by agencies other than SEC; and
6. existence of material internal control weaknesses.

R&W claim that at least four of these six attributes could be interpreted as features of the control environment as defined in the proposed AICPA Statement on Auditing Standards, "The Auditors Responsibility for Assessing Control Risk." They state: "The complexity and variability of expression used by auditors in describing such characteristics as a quality of internal control, management competence, etc, suggests that future empirical work will continue to require the collection and analysis of a large number of possible explanatory variables which may be interrelated."

Waller's (1993) first interest was not in the association between the rate of misstatements and an auditor's risk assessment, but in the question whether risk assessment on the level of assertions (see 3.3.3 figure 1) will lead to improvement for the total risk assessment: that on the level of an account. He states conditions[10] under which assertion oriented risk assessment might lead to improvement of the total risk assessment: (1) varying rate of misstatements over assertions for an account; otherwise there would be no gain in effectiveness from extending the decomposition of risk to the assertion level (2) varying risk assessments by an auditor, over assertions for an account; otherwise the assessments would be informationally redundant (3) positive association between the rate of misstatements and an auditor's risk assessment; otherwise auditors' ability to perform the task would be in doubt.
Waller tested four propositions:

1. There is an association between auditors' IR and CR assessments
2. The rate of detected misstatements varies over assertions, after controlling for DR
3. Auditors' IR and CR assessments vary over assertions for an account
4. There is a positive association between auditors' IR assessments and the rate of detected misstatements, after controlling for CR and DR.

In 215 real audit cases, at the assertion level no confirmation was found of the first proposition; almost all 15 Kendall correlation coefficients that are used in the data analysis were close to zero. The correlations increased to approximately .20 when the cases with CR = "high" were left out of the analysis. So there was a moderate indication that IR and CR are positively related in the cases where the auditor relies to some extent on the controls. This is consistent with the expectation that weak controls produce error prone conditions and inconsistent with the expectation that a high IR will be compensated by strong controls.
The rate of detected misstatements appeared to vary significantly over assertions, for instance for trade accounts payable "completeness" had a rate of 0.115 and "ownership" had a rate of 0.024. Comparable differences were found with inventory and with trade accounts receivable. The differences were all highly significant. Controlling for the detection risk, did not make much difference, because the detection risks only varied between 0.04 and 0.11.
The risk assessments over assertions were found to vary only minimally.

[10] It is remarkable that Waller does not mention that (less) complexity of the assessment task as possibly leading to improvement.

A positive association was found between IR and the rate of detected misstatements, approximately 0.10 (Kendall's correlation) when the three most important assertions were combined; all significant at less than a half percent and for each of these "most important combinations" fully consistent.

Asare & Davidson (1995), Kreutzfeldt & Wallace (1990), Wallace & Kreutzfeldt (1993) and Wright (1994) investigated consistency of risk assessment with the size of errors.

Asare & Davidson (1995) gave a small scale review of studies into risk assessment as a predictor for the error. In three of the studies in their review a positive correlation was found between either contextual or control risk factors and the error found, but there was no explicit risk assessment. In two of these studies the correlation was negligible. A & D performed an own study into the influence of the strength of controls on the error expectation of auditors. They found that strong control procedures lead to predictions of smaller errors. Obviously this finding relates to consistency in risk assessment and not to the relation between controls and errors as such, because the study did not extend to real errors.

Kreutzfeldt & Wallace (1990) tested a total of 75 operational variables for control structure elements by correlating these variables to total error rates and errors at the account level. With respect to total error interrelationships, over one half of the measures tied to the control environments were significantly related to the incidence of error.
Seventy percent of the accounting system measures and hundred percent of the control procedures at an aggregate level were strongly associated with error.
Certain contextual and control dimensions of the company did not relate to errors. Specifically public versus private status, the extent to which management is dominated by one or a few individuals, the existence of significant equity or debt offerings, and safeguarding of investments did not have a demonstrable association with error.
The implications for practice of public accounting are that auditors can measure the effectiveness of control structure of variables in a manner which appears to track in proportion with errors. In other words, clients judged to have more effective controls are observed to have fewer or less severe errors. K&W give the advice to consider inherent risk and context jointly with control structure as they evaluate control risk.
The correlations for occurrence of errors (with control structure) and for size of errors (with control structure) they found, could differ to a substantial extent.

Wallace & Kreutzfeldt (1993) investigated the influence that 5 factors have on the error rate, in 1506 files of completed audits. (1) Management competency, (2) management integrity, (3) company's financial condition, (4) management controls and (5) detailed controls, all appear to have a significant positive relation with the error rate found.

Wright (1994) examined the incidence, impact, direction and cause of detected misstatements as related to the assessed strength of internal controls.
Data on detected errors were gathered from a random, cross-sectional sample of 186 audit engagements. Auditors reported detailed information on 731 detected errors.
The results indicated that as assessed internal controls weakened, the frequency of errors increased and errors were more likely to have an effect on income. Errors were more likely to reflect understatement of assets and liabilities when controls deteriorated, while, when controls were strong, assets and liabilities were more frequently overstated. Further, the causes of errors reflect a greater frequency of "the routine" errors as controls deteriorate, although cut-off errors were relatively common across all control settings. These results suggest different audit strategies are appropriate in response to variations in controls.

Intermediate conclusions with regard to consistency with error rate.
The studies reveal relationship of varying strength between risk assessment and audit opinion and also between risk assessment on specific risk factors and the occurrence or size of errors. (Note that the second relationship does not necessarily imply a relation between risk assessment and error rate)

Discussion
The weak correlation found by Waller between IR and CR, is hard to interpret. The question is whether "in reality", when the IR is high, the controls also "in reality" tend to be of relatively less quality. Or that the correlation is caused by some anchoring mechanism in the assessment itself of the auditor.
Varying rate of misstatements over assertions, found by Waller, combined with hardly varying risk assessments over assertions, indicate a possible improvement of risk assessment. This improvement would entail a valid combination rule of risks assessed over assertions into a risk assessment on the level of the account. (see also Amer et al, 1994 and Srinindi & Vasarhelyi, 1986).
Waller and Wallace & Kreuzfeld aim at a relation with the occurrence of errors. This relation does not necessarily imply a relation with size of error (Kreuzfeld & Wallace, 19990).

Intermediate conclusion of 3.4.2: consistency

Consistency of risk assessment with audit planning, or with that of other auditors, or with the error rate found in the corresponding account, tends to be susceptible to influences, such that consistency can not be relied upon as a stable feature of risk analysis.

3.4.3 Studies on complexity of the object of risk assessment.

Abdolmohammadi & Wright(1987), Tan et al (2002), Van Kuijck (1999), Colbert (1988), Colbert (1989), Bell & Carcello (2000) and Buckless (1989) (among others) deal with complexity.

Abdolmohammadi & Wright(1987) investigated the effects of experience on decision making in auditing. They provided evidence that the experience effect is significant when task complexity is explicitly considered. They reported the results of a series of experiments examining structured, semi-structured, and unstructured tasks where subjects are pooled into two groups: "experienced" (those having reached the staff level where the required monitor skills are developed) and "inexperienced" (lower staff levels or auditing students). Responses to a separate study of 88 partners and managers were used to independently establish the appropriate staff level for each task and complexity. Significant decision differences were found between the experimental groups on each task. A & W state that these results suggest that auditing students or less experienced junior auditors are questionable surrogates for CPA's in complex audit decision settings.

Tan et al (2002) investigated the impact of accountability (degree of responsibility for the outcome of the audit) and knowledge ("of necessary substantive and compliance tests (…) of double entries and financial statement errors") on the performance of an auditor with varying task complexity. They find that for auditors working with high accountability and low knowledge and also for those with low accountability and high knowledge, the performance for increasing task complexity decreases. For auditors with high accountability and high knowledge the performance stays at the same (high) level and for those with low accountability and low knowledge at the same (low) level, when task complexity increases.

Van Kuijck (1999) investigated judgment performance related to experience, education and complexity. His key findings were:

- Judgment performance in terms of the effectiveness of subjects with a university education is better than that of subjects lacking such education. The non-university group made significantly more errors in interpreting accounting records as compared to the other group.
- The university group used significantly more time than the non-university group, but, corrected for errors found, there was no difference in efficiency.
- More experienced auditors did not perform better than auditors who were inexperienced.
- More experienced auditors used more time for their judgments task.
- Task structuring influenced the effectiveness as well as efficiency: both improved under the structured task as compared to the unstructured task. Moreover, there was a tendency that the difference between the routine and complex audits decreased. However, there was no indication that the subjects learned under the structured task.
- Attention was paid to difference in performance of auditors with different educational backgrounds: auditors with a university education under the structured instruction did not perform better than the other auditors. However, these results were based on only five observations.

Van Kuijck investigated a judgment task which is not really aimed at assessing risks. Still the results are interesting, because his judgmental tasks refer to complex situations and so are similar to our risk assessments tasks.

Colbert (1988) examined four inherent risk factors in inventory: (1) rate of turnover of the controller, (2) financing pressure, (3) the complexity of overhead in inventory, and (4) the quality of the personnel responsible for the inventory calculation. Sixty-five practising auditors were included in her study. The results suggest that although all four inherent risk factors were important to auditors, quality of personnel was the most significant in determining the assessed risk.

Colbert (1989) conducted a literature study in an attempt to integrate the findings within various areas on the influence of experience on the quality of judgments. She concluded that in the more structured assessment tasks experience does not show a positive relationship to consensus. But in more complex or relatively unstructured tasks, experience makes a difference in the judgments: the judgments improve.
The findings from studies in internal control support the hypothesis that experience may be vital for complex or unstructured decisions, but not significant for relatively simple or structured judgments. This is consistent with her other finding that, although often in research it is important to classify subjects by experience levels, used years of experience or rank within the firm as surrogate measures for expertise are not appropriate: juniors may be better able to observe inventory than partners, possibly due to their recent experience and detailed instructions from seniors. In other words, experience and expertise are not the same.

Bell & Carcello (2000) did not directly refer to experience of the auditor or complexity of the assessment task. But they did investigate the effect of decomposing the complex assessment task by looking at the determinants for the detection of fraud. In that study they refer to SAS No 82, which makes the auditor responsible for the detection of fraud. Their study used a sample of 77 fraud engagements and 305 non-fraud engagements. In a logistic regression analysis they found six significant risk factors and one interaction term: (1) weak internal control environment, (2) rapid company growth, (3) inadequate or inconsistent relative profitability, (4) management places undue

emphasis on meeting earnings objectives, (5) management lied to the auditors or was overly evasive, (6) the ownership status (public vs private) of the entity and (7) an interaction term between a weak control environment and an aggressive management attitude toward financial reporting. Their logistic model was significantly more accurate than practicing auditors in assessing the risk for the 77 fraud observations. There was not a significant difference between model assessments and those of the practicing auditors for the sample of non-fraud cases.

Buckless (1989) aims at providing information on what factors are involved in the assessment of audit risk and to provide insight into the manner auditors assess audit risk. He conducted two interrelated experiments.
The first experiment was concerned with determining the relative influence of various risk cues on auditors' risk assessments. In this experiment, audit managers were given a list of risk cues and asked to indicate the relative influence of these cues on their risk assessments. The audit managers were assigned to one of two groups. One group evaluated the risk cues with respect to a specific account. The other group evaluated the risk cues with respect to a specific audit objective.
The second experiment was concerned with modelling auditor's subjective assessments of audit risk and examining the effect of the judgment task on auditor's risk assessments. In this experiment, audit managers were asked to evaluate the audit risk of several audit cases.
The case profiles presented varied by manipulations of risk components. The manipulations were selected based on the first experiment. Again, audit managers evaluated the audit cases either with respect to a specific account or with respect to a specific audit objective. The experiment employed a 2 x 2 mixed factorial design.
Buckless found four major points.

- First, audit risk assessments are differentially affected by risk cues.
- Second, auditors combine the risk components in an additive fashion and achieve lower audit risk than suggested by the audit risk model exhibited in the authoritative literature.
- Further, auditors' risk assessments are affected by the judgment task.
- Finally, consensus is higher for risk assessments made with respect to audit objectives (assertions, see 3.3.3) as compared to risk assessments made with respect to accounts.

The second bullet above refers to inadequate aggregation. Aggregation was extensively discussed by Lea et al (1992).

Both Bell & Carcello (2000) and Buckless (1989) deal with complexity by studying the role of various risk factors or determinants of risk assessment. This view is explicitly taken by the studies that are dealt with in the next sub-section.

Intermediate conclusions of 3.4.3: complexity.

Complexity is investigated with respect to experience and some other attributes. 'Experience' appears to have influence on the quality of the assessment task, but not in a straightforward way. Actually sometimes 'expertise' is the better concept to deal with what an auditor learns in practice and by maintaining his education. Experienced auditors are better in dealing with complex situations than novice auditors. But decomposing a complex situation into risk factors and combining them, like Bell & Carcello (2000), in a logistic regression performs better than even experienced auditors do. And accountability moderates the relation between knowledge and complexity.

3.4.4 Studies on tests of control as predictor for misstatements.

In this is subsection we deal with studies that investigate the relation between tests of control and the error rate. As stated in section 2.4, predictive power of the tests of control for the error rate is a necessary condition for these tests to serve as underpinning of risk assessment. Roberts & Wedemeyer (1988), Bell et al. (1998), Blokdijk (2001, 2004), Koning (2002), Blokdijk (2002) and Van Leeuwen & Wallage (2002) deal with this question.

At least four of the determinants (1, 2, 4 and 6, see 3.4.2.3) mentioned by Roberts & Wedemeyer (1988) that appear to have predictive value for the error rate can be assessed by means of tests of control. Therefore it may be expected that tests of control have predictive value for the error rate.

Bell et al. (1998) examined the differential impact of computerisation on common attributes of audit differences. Consistent with prior studies, this study indicated that the majority of audit differences (misstatements) arose due to incorrect computations, differences in management and auditor judgment, faulty initial identification and processing of transactions, and overworked accounting personnel. Likewise, audit differences related to control attributes were usually associated with inadequately skilled personnel, improper or inadequate independent verifications, or improper documents and records; audit differences were readily associated with inadequate controls over assets or records. This study reports additional findings that incorrect manual computations, the recording of exchange documents, incorrect application of internal controls, and inadequate internal controls were more likely to be sources of problems when information systems are computerised than when they are not. Finally, very few of the audit differences in this study were associated in any way with failures in the computerised system. What the study does show is that if errors occur in computerised systems, manual processing and human errors are the most likely to be the cause of the errors. And, consistent with what they found in prior studies, this could be generalised into a proposition that the operation of the system is a greater cause of errors than its design.

Blokdijk (2001, 2004) concluded in a logical analysis that the systems approach, with tests of control as a major part (ISA 400) cannot give definitive evidence on the quality of the annual accounts and is therefore insufficient as a basis for an audit opinion. This means that substantive tests will always be necessary.

Koning (2002) challenged Blokdijk's views: systems assessment, especially by performing tests of control, is a necessary part of risk analysis. Koning gives arguments and refers to the report of among others the Panel (2000) to come to his conclusion that the assessment of the internal controls gives the expected information regarding the quality of the annual accounts.

In reaction Blokdijk (2002) states that the Panel (2000) does not show the effectiveness of tests of control where it regards the truth and fairness of the accounts. It would be better to make the tests of control part of the guiding activities for an audit, similar to the use of analytical procedures. In a postscript, Koning states that especially in an automated environment it can be very effective with regard to the truth and fairness of the accounts to test the proper operation of controls.

Van Leeuwen & Wallage (2002) give an introduction into the business process analysis. It extends the usual risk assessments in the audit risk model to the assessments derived from analysis of the business processes and the context in which the business is operating. It claims that the extension implies more assurance with regard to the

audit opinion. It associates the proper design and operation of the internal controls with a low occurrence risk (p 87).

Intermediate conclusions of 3.4.4: validating power of tests of control.

Only a few studies are included in this sub-section. They indicate predictive power of systems characteristics for misstatements, but far from exclude human errors as a cause for failures. The cited Dutch discussion on the value of tests of control as sufficient evidence on the quality of the annual accounts, adds to the relevance of investigating the predictive power of the operation of the system of controls for the existence and size of errors, as performed in this thesis.

3.5 Discussion

Not all expectations given in section 3.2 are covered by the literature discussed in this chapter. The expectations with regard to representativeness were neither clearly confirmed, nor rejected, when overweighting of sample results or of prior probabilities was concerned. Predictability or regression effects were not studied in the given literature. The expectations with respect to the availability and anchoring heuristic were confirmed. So the presented studies reveal a lot about the heuristics and biases in the assessment of auditors. In general, the studies show that this paradigm is a fruitful one. In most cases the heuristics appear to be active. But, encouraging for the profession, proper training or experience can mitigate the related biases. And experience often is sufficient as that proper training, although care must be taken for the possibility that the experience is only of value when it also implies expertise with respect to the audit task. Expertise is then to be seen both from the viewpoint of auditing and from knowledge of the kind of organisation or industry whose accounts are audited.

Direct validation of risk assessment in practical situations forms only a minority in all the studies we discussed[11] (see: Waller, 1994, Asare & Davidson, 1995, Kreuzfeldt & Wallace, 1990, Wallace & Kreuzfeldt, 1993, Wright, 1994, Bell & Carcello, 2000) The studies mentioned in the heuristics and biases paradigm do not use real-life situations in order to validate the assessed the risks on the real incidence or size of errors. The same applies to the other types of study. This is not compensated by the fact that in many of the studies some quantitative risk or probability is specified which has to be assessed in some way or another by the experimental subjects. For such a compensation it would have to be sure that the primitive risk assessment tasks on primitive events, however defined, will be done in a valid way. And even if this were the case, hardly any literature is found on this topic. Moreover, what numerical value is associated with a 'high risk', 10 %, 20 %, 40% or even more? Concern for the answer to this question is directly implied by the conclusion of 3.4.2.1.

So we can extend our conclusion that the heuristics and biases paradigm is interesting, with the conclusion that research based on it does not give conclusive information on the validity of risk assessment in practice. Similar conclusions can be drawn on the consistency studies: our review shows very interesting things on the conditions under which consistency in the assessments will be seen and/or will indicate more quality of risk assessment. Also in these cases, however, no validation is performed on the error rate or on an estimated audit risk, except for a couple of studies.

[11] Maybe our strategy of selecting studies is the cause for this: include the articles with titles that refer to risk assessment. But it is not very probable that from a body of literature with a reasonable portion of this "validation-type" of studies, only so few studies would be found.

So here is a gap to be filled. In this thesis this work will be started.

Various studies go into the distinction between risk assessment on the assertion level and risk assessment on the account level. They suggest that risk assessment at the assertion level has more quality than risk assessment on the account level, but as Kinney (1992) notes: the ARM fails to reflect the risk of joint misstatements in two or more assertions. And although Srinidi & Vasarhelyi (1985) give suggestions to solve this problem of aggregation, in practice they are not followed, at least in our own experience in governmental audit practice and in the practice of some private audit firms.

The complexity of the assessment task is dealt with in several ways. One way is to try and find determinants of the error rate in a direct sense, thereby circumventing the problem of risk assessment. Still for these determinants a major assessment is often necessary. The other way is to try and find determinants for risk assessment. In both types of this research these determinants are found. In other words: they show that it helps to decompose the assessment task in an assessment task on smaller scaled problems.

The studies on consistency of risk assessment and audit design are the most consistent in showing a positive relationship. Of course, the practitioner that operates in accordance with the auditing standards, should be expected to show this consistency in his assessments and choices in audit design.

3.6 How to get insight in the 'real risk'

We have made it our aim (probably with approval of Reimers et al, 1993) to get insight in the 'real risk' of a material error, more precisely, how far this can be assessed by the auditor by means of risk analysis according to the ARM. We have found many studies on the quality of risk assessment, but only some in real life situations, with the error rate or related measure as a criterion (see 3.4.2.3). We summarise:

- consensus appeared not to have consistent positive influence on the accuracy of an error estimate (Davis et al, 2000),
- attributes of the control environment appeared to have predictive qualities for the error rate (Roberts & Wedemeyer, 1988; they suggest research into a large number of explanatory variables),
- a Kendall correlation of .10 is found between IR and the rate of detected misstatements (Waller, 1993)
- some weak relation between contextual or control risk factors was established by Asare & Davidson (1995); they did not investigate risk assessment as such
- operational variables regarding the controls structure were found to have strong correlations with the error rate (Kreuzfeldt & Wallace, 1990)
- both factors of inherent risk and of control risk have a positive correlation with the error rate (Wallace & Kreuzfeldt, 1993)
- strength of internal controls and incidence and impact of errors are positively related (Wright 1994)

Studies in this summary sometimes show a clear correlation, between risk factors and error rate, sometimes a weak correlation, between IR and rate of misstatements. We have to realise that investigating risk factors is something different from investigating risk assessment. Moreover the rate of misstatements (relative number of misstated transactions) is different from their size. So there are two reasons for completing the quoted research by investigating the validity of risk assessment with respect to the error rate (as relative number of monetary units) and related measures:

- can we extend the positive correlation for IR and rate of misstatements as Waller found, into one for OR and error rate?
- do the high correlations between risk factors and the error rate hold, when instead of risk factors, risk assessment is taken as the predictor.

Taking into consideration what we found in chapter 2 with respect to appropriate validation criteria, the following question emerges from this overview and discussion, and will be taken as leading question for our research:
1. Has the assessed risk by an auditor, predictive value for the error rate or an empirical estimate of the audit risk, such as the sampling risk?
Naturally, we are interested in factors that influence the strength of the relation formulated in this question. This is to say that we are looking for moderator variables, as in the next question.
2. Can we find moderator variables, influencing the strength of the relation meant in the previous question?

There is a striking difference in the strength of the relations found by Waller and most of the references investigating separate risk factors. This may be explained by a difference in complexity: the risk assessment task is complex compared to the assessment of one factor. In this chapter we have seen that in more complex situations, heuristics like anchoring, availability and representativeness are more likely to lead to corresponding biases. Several authors (Colbert, 1989, Bell & Carcello, 2000, Buckless, 1989, Abdolmohammadi & Wright, 1987) actually showed that complexity has influence, and also that it sometimes can be dealt with, a) by making assessments at the assertion level (for instance Waller, 1993, Buckless, 1989), b) by excluding judgment as far as possible: logistic regression on directly measurable risk factors performs better than judgment of auditors (Bell & Carcello, 2000), c) by decomposing the assessment needed into part- assessments (Roberts & Wedemeyer, 1988, Wallace & Kreutzfeldt, 1993).

In this study we investigated the possibility to improve risk assessment by decomposition of the assessment task into a set of subtasks, each of which would be less complicated. Each subtask consisted of giving an assessment regarding the risk of a material error on a ‘risk indicator’. The question then is:

3. Can risk assessment be improved by decomposing the task into risk components or risk indicators?

We had two related reasons for aiming at the tests of control (which we will call ‘system tests’ in chapter 9):
(1) these are necessary as an underpinning of risk analysis, when an assessment ‘low’ is to lead to a decrease in the extent of substantive testing (ISA 200, par.21, HCDAD 1997) and
(2) we concluded (section 2.4) that reduction of the extent of substantive testing can only be justified if tests of control have predictive power for the error rate.
So we decided to investigate a fourth question, a decision with an extra justification from the discussion cited in 3.4.4:

4. Are tests of control fit for underpinning risk assessment?

Including this fourth question was also caused by a very practical reason: we were granted access to a large set of data giving an opportunity to answer this question.

By concentrating on these questions, our research aims at validating risk assessment directly on the error rate or an estimate of the audit risk, thereby circumventing the

possibility that a risk assessment which is fully consistent, consensual and not suffering from ‘Tversky and Kahneman bias’, might still miss the “real risk”, because the assessment of the risks associated with the primitive events in the administrative process happened to be biased, the possibility that was discussed in section 3.5.

In chapter 5 we will discuss the research design in which we tried to find answers to these questions.

To conclude this chapter we give a table with a summary in keywords of the things we found in the cited studies.

Table 3.3: An overview of the findings in literature.

Authors	Framing effect found?	Anchoring effect found?	Recency or availability effect found?	Represen-tativeness effect found?	Positive experience effect ?
Johnson et al (1991)	yes, 1)	yes 2)			Yes/no 3)
Emby (1994)	Yes, but only in interaction with recency		Yes, but only in interaction with framing		
O'Donnel & Schultz (2005)		yes			
Wilks (2002)		yes			
Butler (1986)		Yes 4)			
Joyce & Biddle (1981a)		Yes 5)	Yes caused by order		Yes: more experience => less anchoring
Krug Nelson (1995)		Yes			
Smith & Kida (1991)		Yes: disjunctive, conjunctive			
Tan (1995)		Yes 2) (consistency)			
Bar Hillel (1979)				Yes	
Joyce & Biddle (1981b)				Yes	
Krug Nelson (1995)				Yes	
Smith & Kida (1991)				Yes	Yes
Messier et al (2001)				Yes	
Burgstahler et al 2000)				Yes	No and yes
Ashton & Kennedy (2002)			yes, but disappears by self-review		
Hall et al (2001)			yes		
Cushing & Ahlawat 1996)			Yes, but disappears by documenting		

1. in case of fraud: independent of auditor experience, dependent of knowledge of business
2. anchoring in the form of consistency: this is a tendency to consider the same aspects in repeat engagements
3. knowledge of the industry caused better performance in discovering fraud; experience as an auditor did not; experience as an auditor caused better performance to detect financial misstatement
   framing effects were less with greater expertise, either as an auditor, or as expert in the business
4. on anchor: allowed audit risk
5. on plain anchor; no difference conjunctive / disjunctive

Table 3.3 (continued): An overview of the findings in literature

Authors	Is audit design consistent with risk assessment?	consensus with respect to risk assessment?	Risk assessment consistent with error rate?	Positive experience effect?
Basu & Wright (1997)	Yes			
Dusenbury et al (1996)	No at the account level, yes at the assertion level			
Mock & Wright (1995)	no			
Joyce (1976)		No but 1)		
Gaumnitz et al (1982)	Yes 2)	Yes 2)		
Srinindi & Vasarhelyi (1986)	No	No/yes 3)		
Houston et al (1999)	Yes, conditionally		Yes, varying for errors or irregularities	
Elder & Allen (2003)	yes			
Stone & Dilla (1994)		Yes 4)		Yes 4)
Reimers et al (1993)		yes / no 5)		
Amer et al (1994)		No		
Trotman & Wood (1991)	Yes 6)	Yes 6)		
Davis et al (2000)		Consensus(=> accuracy) 7)		Yes 7)
Roberts & Wedemeyer (1988)			Yes 8)	
Waller (1993)			Yes, moderately 9)	

1. there is consensus between auditors of the same firm, much more than between those of distinct firm.
2. possibly because a quantified assessment of the strength of internal controls is asked for (where the planned work served as the assessment of this strength in Joyce 1976)
3. dependent on aggregation rule and primitive assessment
4. when risk classification is in numbers and subjects are experienced auditors; else no consensus
5. high consensus when risk categories are linguistic; much lower consensus when the categories are numerical
6. based on a meta analysis of 17 studies
7. the relation “consensus => accuracy” has varying strength over types of industry. Experience appears to have a positive influence on the strength of the relation between consensus and accuracy
8. this consistency regards the relation of six ‘general attributes and the error rate; somewhat different from proper risk assessment, but similar enough as to be of interest for this thesis
9. Waller rather investigates the relation of risk assessment and the audit opinion resulting from the audit

Table 3.3 (continued): An overview of the findings in literature

Authors	Consistency with others: consensus	Consistency with error rate	Complexity	Experience
Asare & Davidson 1995		Yes, but not stable		
Wallace & Kreutzfeldt 1993		Yes 1)		
Wright 1994		Yes, but not consistent		
Abdolmohammadi & Wright 1987			Yes	Yes, dependent on complexity 2)
Tan et al 2002			yes	yes, dependent on knowledge
Van Kuijck 1990			Yes 3)	
Colbert 1988			Yes 4)	
Colbert 1989			Yes 2)	Yes, dependent on complexity 2)
Bell & Carcello 2000			Can be circumvented by decomposition	
Buckless 1989	High on assertion level ; low on account level		Leads to inconsistent combination rules	

1. 5 factors (overlap of 3 with Roberts & Wedemeyer, 1998) appear to have a significant positive relation with the error rate
2. experience makes a difference in complex situations; not in relatively uncomplicated situations
3. experience, education and task complexity/ structuring affect assessment
4. but quality of personnel had more influence on the assessment of IR

Chapter 4: The Sampling Risk[12]

One of our validation criteria is the sampling risk. We calculated this risk as a probability in the beta-distribution. In this chapter we show (in a technical way) that this approach has good properties. A by-product of this chapter is a method for calculation of the upper error limit that is much simpler than the usual Stringerbound, with an extra advantage that it gives the sampling risk and is less conservative.

4.1 Introduction

In this chapter we will introduce a method for calculating the sampling risk, when we only have the sample size and the estimated error rate at our disposal. In this section it is made plausible that this method may lead to useful results, in the second section a simulation study is designed to substantiate this plausibility. In section 4.3 we give results of this simulation and section 4.4 discusses their generalisability.

In chapter 2, 2.2.6 we came to the conclusion that the sampling risk $P(p>M|D)$ is the generic measure for the occurrence risk. This expression is to be read as "the probability (P) that the error rate p, associated with the quality of the administrative process, exceeds the level of materiality M, given the data (D)".

Now the problem is which probability law applies to this sampling risk. In 2.2.6 we saw that P(p>M|D) is calculated with the help of

$$f(p|D)=(f(D|p)f(p))/f(D) \qquad (1).$$

In this formula f(D|p), the likelihood, will be dependent on the sampling design. In auditing the so-called “monetary unit sampling (MUS)” is the most widely used. In MUS the distinct monetary units (MU’s, e.g. dollars, guilders, euros) are seen as the unit of sampling. So the population from which the sample is taken is the collection of all booked MU’s. Every single MU is selected with the same probability (1/N, when there are N MU’s). The question is whether this MU is in error or not.

In the case of the "all or nothing approach" - "the my euro right or wrong approach" – every “*book-value*” (an entry in the account containing the value of the transaction to which the book-value corresponds) is subdivided in the MU’s that are correct and those that are not. The selected MU is in error when it comes from the part of the corresponding book-value that was in error. In this approach for every MU it is defined whether it is in error or not. Say that totalled over the whole account a fraction of p is in error. Now when a sample of n MU’s is selected with replacement, k, the number of MU’s in error in the sample, will have a binomial distribution with parameters (n,p). The hypergeometric (for sampling without replacement), or the Poisson distribution (when p is small and n is large) can also be used. In the evaluation of the sampling risk, the so called beta distribution also plays a role, as we soon will see.

The fraction MU’s of a book-value that is in error is called the *‘taint’* or *‘tainting’* symbolized by T. When an approach with taintings is used, or when only the most likely error (MLE) and the sample size are available, as in most cases in the study, the probability law is unknown. The aim of this chapter is to investigate a possible probability model for this situation. We start with a heuristic argument to show where

[12] This chapter gives a justification of the way we calculated the sampling risk in chapters 6 and 8. These chapters can be read without having read chapter 4.

the idea for the solution comes from.

The hypergeometric, the binomial and the Poisson distribution all apply to integer valued outcomes. But the outcomes of most of the cases that are in the study are not given in the number of errors found in the audit (sample), but as an estimate of the error rate in the population. And at its best, only by exception, this estimate can be converted into an integer number of errors in the sample of the relevant case, which would be a necessary condition for the applicability of one of the three probability laws.

This means that we have to find another probability law in order to evaluate the sampling risk that is given with the outcome in the cases that were in the study. We will give an intuitive argument that the beta distribution is fit for this purpose.

The beta distribution has the following density:

$$f(p)= (1/ B(a,b))p^{a-1}(1-p)^{b-1}; 0<=p<=1; a,b>0$$

In this density, B(a,b) is the beta function, a factor dependent on a and b, which causes the probability for all p's ($(1/B(a,b))\int_0^1 p^{a-1}(1-p)^{b-1}dp$) to be equal to 1.

A well-known theorem in Bayesian statistics states the following (see Lee, 1997):

Property 4.1:
Let p (0<=p<=1) have a beta prior distribution with parameters a_1 and b_1. Let $\underline{k}$ have a binomial distribution with parameters n and p. Then if $\underline{k}= k_0$ successes (errors) are found in a sample from this distribution, the posterior distribution for p is also beta, with parameters $a_2 = a_1+k_0$, $b_2 = b_1+(n-k_0)$.

This property implies that the beta distribution is similar to the binomial distribution; in Bayesian statistics the distributions are called conjugate. For our purposes the beta distribution has a convenient property: it uses broken (or rather real valued) parameters, a and b. These parameters play a role which is similar to the number of successes (a has this role) and the sample size (a+b has this role) in the binomial distribution.
We observe that in property 4.1 k_0 is integer valued, but a_2 and b_2 are real valued. The binomial law has no meaning for a broken k_0, but the beta posterior (and also the beta prior) is defined for every value of a and b (both > 0), so also for every broken or real value. Now a value for a_2 of for instance 2,3 in the posterior beta distribution can result from $a_1 = 2$ and $k_0 = 0,3$ (if this value for k were possible). But it can also result from $a_1 = 1,3$ and $k_0 = 1$. So, for the beta distribution it does not matter whether k is integer valued or not.
We can also observe that the beta distribution is continuous in a and b, which means that it also takes on values for all a and b between the integer values and so, in a sense, gives a way for interpolation between these values. This leads to the idea that it makes sense to model the sampling risk with the beta distribution as follows.

Beta-model
Let r be the rate of error in a sample of size n, let M be the level of materiality, let p be the error rate in the account under audit and let B be the beta distribution with parameters a = 1+r*n, b = 1+ (n-r*n), then modelling with B of the probability that p is larger than M gives a good approximation to the real probabilities.

In the beta-model r*n takes the place of k in property 1. The specified beta distribution can be seen as the posterior distribution for p, as the result of a uniform prior

(beta(1,1)) and a likelihood for r, which is not known, but is similar to the binomial distribution. Therefore this distribution will be referred to as the 'beta posterior'. The beta-model has some properties that add to its plausibility.

The first one regards the mode of the posterior distribution, formulated in property 4.2.

Property 4.2:
The mode of a beta distribution with parameters a and b is equal to (a -1)/(a + b-2).

For the beta distribution of the beta-model this means that its mode is equal to r*n/(r*n+(1-r)n) = r. This is consistent with the most likely error, the estimate for the rate of errors in the population, which is also r. It also gives an extra justification for the model

The second consequence regards the fact that for the binomial case the distribution of the beta-model is the same as the one in property 1.

Now these two consequences are far from a mathematical proof of the beta-model. In fact, a proof of this beta-model is impossible, because a fit of practical data to a mathematical model never can be proven; it can only be shown. But we may hope that a mathematical form can be found for data modelled in a plausible way (such as the binomial model for sampling with replacement), combined with a prior beta distribution. We did not succeed in finding such a model with related proof.

Therefore we started a sub-project with Dr Karma Dajani, lecturer at the Department of Mathematics at the University of Utrecht and Bianca Snel, a student in mathematics at the same University. In this project we conducted a simulation study and tried to find a mathematical expression for the posterior distribution if we assumed a beta prior for p, the error rate in the population. The remainder of this chapter will give a summary of the results, which are fully reported in Snel (2002).

4.2 The design of the simulation.

The distribution of the error is not known and the way the sample was taken varies over cases. Therefore it was convenient to investigate (and therefore generate) distributions of errors (in the form of taints) that generally are assumed to be valid models of real error distributions. These models have the form of a mixture of two distributions. In addition we also investigated an unmixed model for the distribution of the taints. It was done both in a classical and in a Bayesian setting.
In the sequel we will use a third property of the beta distribution.

Property 4.3:
The expectation of a beta distribution with parameters a and b is equal to a/(a+b).

For the simulation, populations of book-values were generated. In the book-values errors were inserted, based on a variety of distributions. From these tainted book-values a sample was taken in which the relevant statistic: the average taint as the point estimate for p and a confidence limit was calculated and given the outcome of the statistic two values were determined:
1. that for the probability of the parameter exceeding some critical value (the materiality), given a prior distribution and the value of the statistic;
2. based on the sampling distribution, that for an upper confidence bound, which was compared to the standard Stringer bound.

To be more precise, the following steps were taken in the simulation study.

1. Three times N accounts AC_i (i=1,…,N) were generated, each consisting of A book-values. In the study N was mostly taken 100000, and A was taken 10000. The size of the book-values was taken from the exponential distribution with average equal to 1.

2. In each account AC_i (i=1,..,N) errors (taints) were inserted, such that the error rate p_i (the rate of MU's in error) in account AC_i is equal to i/N, so p_i=i/N. This was done in three different ways, one way per N accounts, labelled as the beta likelihood, Mix1 and Mix2, as follows:

2.1 taints were inserted distributed according to the beta likelihood with parameters (p/(1-p),1); as a consequence of property 4.3 this distribution has the expectation of (p/(1-p))/(p/(1-p)+1)=p [13]

2.2 taints were inserted distributed according to a mixture of two distributions, g and h as follows:

$$f(T_i|p) = q.g(T_i|p)+(1-q)h(T_i|p).$$

Here

T_i is the variable representing the values T_{im} (m=1,…,A) for the taints in the i'th account AC_i;

g is the density of the beta distribution with parameters (*bp/(q-p), b);* (b will be varied to render two sets of mixtures Mix1 and Mix2)

$h(T_i|p) = 1$ *for* $T_i = 0$ *and* $h(T_i|p) = 0$ otherwise;

q is the fraction of book-values with non-zero taints.

This mixture implies the taints in the book-values to be distributed as follows: with probability 1-q they take on the value 0; with probability q they take on the value of the specified beta distribution g.

Note that the distributions of 2.1 and 2.2 apply to T, the fraction of MU's in error in a book-value, whereas the beta prior and the beta posterior in this simulation apply to p, the fraction of MU's in error in the total account.

As a consequence of property 3 the beta distribution g has an expectation

$$E(g) = (bp/(q-p))/(bp/(q-p)+b)=p/q.$$

So the mixture $f(T_i|p)$ has expectation:

$$E\ f(T_i|p) = q.p/q+(1-q).0=p.$$

In the simulation in this mixture the next values were taken for q and b: q= 0.05, q=0.10, q=0.15; b=1 (Mix 1), same values for q and b=0.5 (Mix 2). Taking b=0.5 causes the beta-distribution to have more mass close to 0 than the distribution with b=1 (see also 4.4.2, 3rd point);

2.1 and 2.2 have as consequence that for one round in the simulation effectively 300,000 accounts were generated (the beta likelihood, Mix 1 and Mix 2, each with an error rate of p, which of course varied per 3 accounts and for the Mixes had q as a maximum. The combination of values for b and q implies 6 rounds of simulation.

3. From each AC_i a sample S_i of n book-values was taken, with a probability proportional to the size of the book-values (PPS sampling). (done for 3 times N accounts AC_i)

4. In each selected sample S_i the average $\check{T}_i$ of the taints T_{im} was calculated : $\check{T}_i=\Sigma_m T_{im}/n$. This was done for the 3 types of distributions described in 2.1 and 2.2.

5. Each $\check{T}_i$ is associated with the p_i of the account AC_i from which the sample S_i was

[13] We hoped to find a mathematical expression for the distribution of the average taint in a sample from this distribution, but we did not succeed. As a second best approach we investigated the properties of this distribution by including it in our simulation study.

taken (for 3 times N accounts AC_i).

6. The $\check{T}_i$ are ordered with respect to their size and subdivided in classes C_k (k=1,..,1000) of size 100, by grouping every 100 successive values; so C_1 consists of the 100 smallest $\check{T}$, C_2 of the next 100 smallest $\check{T}$ of the 99,900 remaining $\check{T}$, and so on.

7. In each C_k the statistic $P(p>M|\ \check{T})$ is calculated, for a level of materiality M. This expression can be read as: "the probability that the error rate in the population is larger than the level of materiality, given the average taint in the sample." This "probability" is not a real probability, but an estimate of the relevant chance, simply calculated by counting the number of p_i in class C_k that exceed M and dividing it by 100. It is done for the 3 types of distributions mentioned in 2, above. Also for each $\check{T}_i$ this "probability" is calculated by means of the beta posterior distribution as it was formulated in the beta-model of the previous section.

Remark: On the average the T_i in a class C_k do not differ more than 1/1000, so that the variability of the $\check{T}_i$ in a class C_k is negligible compared to the random variation that will be present in the associated p_i.

8. In each C_k for each of the 3 distributions (Mix 1, Mix 2 and the beta likelihood for T) and also for the beta posterior upper confidence bounds UB, with a level of confidence of 1-β, are calculated in the following way:
UB is the value for which $P(p>UB|\ \check{T})=\beta$, making UB the 1-$\beta$ quantile of the p_i's that are connected to the $\check{T}_i$ in class C_k. In the expression $P(p>UB|\ \check{T})=\beta$ the median of the $\check{T}_i$ in class C_k is taken for $\check{T}$. Now this UB is not an upper confidence bound in the classical statistical sense. But its value is very close to that of an upper confidence bound in the classical sense (see for instance Novick and Jackson (1974, pp120,121). The relevant probability in classical statistics attaches to the method by which the confidence interval was derived (and is called 'confidence level'), in Bayesian statistics it attaches to the particular interval (0, UB) found.
Next to these UB's, also the Stringer bound, SB, is calculated for a confidence level of 1- β. This is done on the basis of the actual taints found in the simulated samples from the Mix1 distribution, in the following way.
For each of the samples in class C_k the taints are known; the SB is calculated from them. This gives 100 SB's in every Ck.
These 100 SB's are averaged and this average is used as the Stringer bound associated with the average of $\check{T}$ in the class C_k.

In the end the simulation produces 1000 classes Ck (k=1,...,1000) and associated with each Ck:
100 values for $\check{T}_i$,
100 values for p_i,
4 values for $P(p>M|\ \check{T})$, estimated for the two mixture distributions and the beta likelihood for T and calculated for the beta posterior,
4 values for an upper confidence bound UB for the same 4 distributions, at a confidence level of 1-β, plus, for the sake of comparison a Stringer bound SB for the same confidence level 1-β.

Now in simulations β was taken 5%, the materiality M was also taken 5%, the sample size n was taken 50. In some simulation rounds, the simulation was only done for the most relevant possible values for p_i: those up to 10% or 15%, depending on the value chosen for q, thereby reducing the time needed for a round. The relevant numbers N=100000 and A=10000 could then be reduced up to N=10000 and A=1000, varying over the rounds.

Just to remember, the simulation study is meant to establish the validity of the beta posterior distribution as a model to calculate the sampling risks that are associated with the cases in the study. When the probabilities calculated in step 7 are equal to or very close to that computed with the beta posterior, the validity is confirmed. Because the estimated probabilities for the beta likelihood and the two mixes are a (close) estimate of the relevant probability in these distributions, so they can serve as a calibration for the corresponding result with the beta posterior in the same situation. So to the extent that the 3 distributions we used cover the distributions that an auditor meets in practice, the probabilities calculated in step 7 are sufficient to establish the validity of the beta posterior.

However, we thought it useful to also look at this validity from a different perspective, namely from that of the confidence bound. With the Stringer bound we have a kind of benchmark for confidence bounds in auditing, so when we compare the Stringer bound with the upper confidence bounds calculated in step 8, we have another means of validating the beta posterior distribution.

4.3 The validity of the beta posterior.

In this section will show the results of our simulation study. In this study the standard level of materiality M = 5% was adopted. Variations were implemented in the mixture distribution: 95%, 90% and 85% were adopted as the rate of error-free MU's (1-q) in the population AC_i.

We show results, given by means of two graphs in figures 4.1 and 4.2, that are typical for the outcomes. In this round the following parameter values for the simulation were adopted: N=100000 A = 10000 (figure 4.1) N=10000 A = 1000 (figure 4.2) n = 50 M = .05 q = .1, prior distribution for p ~beta (1,1) (both figures)
The abscissa in the figures represents T

Figure 4.1.

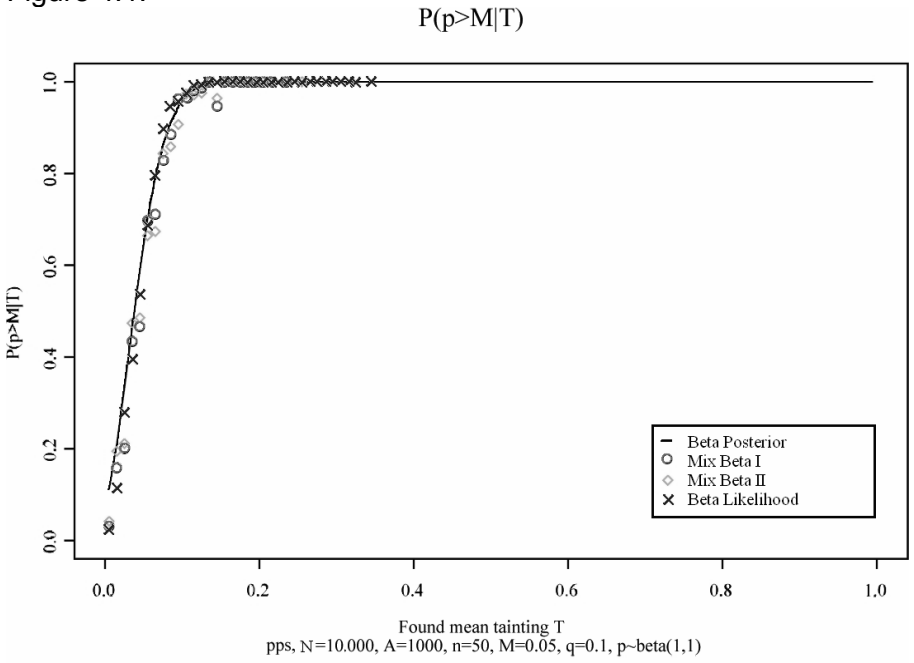


Figure 4.2

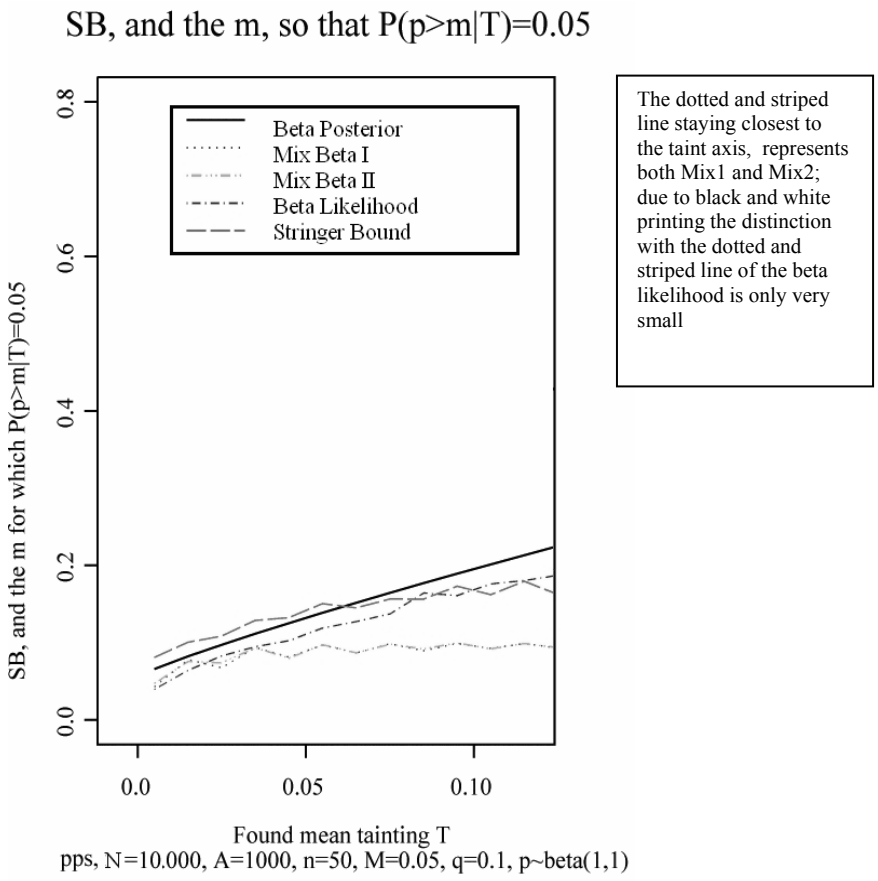


Figure 4.1 shows that the probabilities for p of exceeding M, are very close to each other on almost the whole range of the mean taintings found. Only for values close to zero the upper-tail probabilities diverge. The probabilities calculated for the beta posterior, the solid line, show what can be expected: when very small taints are found, the probability of exceeding M is not getting too close to zero. From the regular audit samples for instance of size 50, we know that when the real error rate in the audit object equals 5%, the probability of finding zero errors is 8%. This is reflected in figure 4.2: here the beta posterior shows an upper-tail probability for M of about 11%. The difference with the 8% binomial probability is partly explained by the fact that the smallest mean taint for which the upper-tail probability of M is calculated in the simulation study, is somewhat larger than zero.
We had to find an explanation for the fact that the three distributions that were simulated show much smaller values for these upper-tail probabilities. We had a suspicion that it might result from a fallacy in the simulation: if for small values of p the taints generated would show insufficient variability, then the upper-tail probability of M would become too small. Some small extra simulations confirmed our suspicion. These extra simulations also showed that this lack of variability only influences the outcomes for mean taints very close to zero, because for the larger values of $\check{T}$ the $p_i|\check{T}$, the values of p, given the mean taint $\check{T}$ have an appropriate spread around M. This also explains why for the larger values of the mean taint the upper-tail probabilities are a little smaller than those of the beta posterior.
The extra simulations also showed that the lack of spread is only small; though large enough to mitigate the upper-tail probability of M, small enough to hardly influence the 95% quantile of $p|\check{T}$, which quantile is the 95% UB (upper confidence bound, upper error limit in auditors' language) that result from the simulations.

Figure 4.2 shows the Stringer bound (SB) and upper error limits calculated as

explained in section 4.1. As can be seen, for mean taints up to 7 percent, the solid line of the beta posterior distribution lies between the Stringer bound (which appears to be more conservative) and the upper error limits that result from the beta likelihood and the two mixed beta distributions. For the assumed distributions of the taints, these upper error limits may be assumed to give a valid confidence bound for p given Ť. Figure 4.2 only shows the part of the graph that is informative for our beta-model, for our beta-model applies to values of Ť that virtually all are smaller than 10% and mostly much smaller or zero. Moreover figure 4.1 shows that for values of Ť larger than ca 10% the upper-tail probabilities all are getting (much) larger than 0.45, which will be the lower boundary of the largest class for the sampling risk in the classification we will use in our data analysis in chapters 6, 7, 8.

Discussion of the results shown in figure 4.1 and figure 4.2.
The results in figure 4.1 and 4.2 are almost completely consistent, with regard to the quality of the beta posterior distribution as it was expected when adopting the beta-model:

- the upper-tail probabilities of M appear to be very close to that of the beta posterior on almost the whole domain of the mean taints; if there is difference the beta posterior is somewhat larger;
- the upper error limits (upper bounds, UB's) of figure 4.2 show a consistent picture: the beta posterior UB is a little larger than the UB's from the two mixture distributions and the beta likelihood. As the UB's from these three distributions may be assumed to be correct for these distributions (this was the way they were derived in the simulation), the beta posterior UB appears to be somewhat conservative.
- Moreover at the most relevant range, that of average taints smaller than ca 5%, the beta posterior bound is sharper (less conservative) than the Stringer bound. This makes the beta posterior bound a good candidate for replacing the Stringer bound

Only the upper-tail probabilities for mean taints very close to zero are a little puzzling, but as explained above these results have been explained by extra simulations, which support the validity of the results.
It can be concluded that the beta posterior gives probabilities that are valid for the distributions that were assumed in this round of the simulation study.

So far we discussed the results for mean taints not greater than 7%. These are the results that would be relevant in case of a real evaluation of audit results and forming of audit opinion as long as the chosen level of materiality does not exceed 7%.

The way the probabilities behave for mean taints larger than 10 percent, in practical situations is not very relevant. This also applies to the interpretation of the results given in figure 4.2. Here we see that the beta posterior bound just grows with the mean taint. This is a direct consequence of this upper error limit just being calculated on the basis of the beta-model: in the simulation the beta posterior bound is just calculated for a mean taint, regardless of its possible existence.

The two mixes show an upper error limit that grows with the mean taints, up to a level that is equal to the value of q, which is 0.1 in this round of simulation. This is necessary of course, because p's larger than 0.1 are impossible with these specifications.

The simulations were also done for mixture rates of 5% and 15%, and for a level of materiality equal to 2.5%. The results were very similar to the ones discussed above. It means that for relevant levels of materiality and for relevant mixture rates of errors in

the transactions of an account, the beta posterior is an appropriate model to evaluate the probabilities associated with the outcomes of the samples in our study.

4.4 Conclusion and Generalisation

4.4.1 Conclusion

For error distributions in accounts like the ones simulated in the current simulation study, the beta posterior upper error limit is better than the Stringer bound: it is a little more conservative than bounds that may be assumed to be valid, and sharper than the Stringer bound.

In this study we will use the probabilities that are calculated with the help of the beta posterior distribution in various kinds of correlation studies some of which with a classification of the sampling risk. Here the absolute value is less important than the relative value. It can be concluded that the beta posterior is fit as a model for the calculation of these sampling risks.

4.4.2 Generalisation

A few remarks can be made on the generalisability of the results to practical situations.
1. The use of the beta posterior as formulated in the beta-model, is only dependent on a proper estimation of the error rate in the population and the sample size. This estimation must be done in accordance with the sample design. But given the estimate of the error rate, the beta-posterior is not dependent on the sample design that made this estimation possible.
2. In the simulation study an exponential distribution was used for the size of the book-values. The skewness of such a distribution is like that of many accounting populations (see for instance Willemsen, 1996). In PPS-sampling the size of the book-values affects the outcome. Especially when the account contains items of an extremely large size, the results might be influenced. But in auditing practice these extremely large book-values are isolated and audited separately from the rest of the account. And then for the rest of the account it is more likely that something like the exponential distribution applies. But we did not investigate the properties of the beta posterior in the light of other distributions for the size of the book-values.
3. In practice distributions of taints occur that show a point mass in 1. The simulation studies of for instance Van et. al. (1996) show that this can strongly affect the results for probabilities associated with values of Stringer bound or other relevant statistics. See also Tamura et. al. (1989). In our study we tried to simulate this reality to a certain extent by also using the value .5 for the b-parameter of the beta distribution as this causes this distribution to have a large mass close to 1. (But this is not the same as a point mass in 1).

4.4.3 A welcome side effect

The simulation study gives a strong indication that the beta posterior is a better evaluation instrument than that provided by the Stringer bound. It has three advantages over this bound:

1. The computations are much simpler
2. Our study shows that it is less conservative than the Stringer bound (at the same time having sufficient reliability.)
3. Relevant probabilities, like the upper tail probability, can easily be calculated, while probabilities for the Stringer bound, given a value for this bound, cannot be calculated directly.

So, provided that the validity of the beta-model can be given a firmer empirical basis than provided by our study, we have given perspective on a convenient evaluation method for substantive test. A simulation study that varies on the distribution of the size of the book-values and also introduces point masses for taints in 1 can give this firmer basis.

Our result can also be seen as a confirmation of the quality of the Stringer bound (including its conservatism).

Chapter 5: Design of the Research

Our research was initiated because of a need to learn more about the 'real risk'. Therefore we collected real life data concerning real life risks in a first study. The outcome of this study and availability of new real life data determined the content of a second and a third study. Their design is also discussed in this chapter.

5.1 Introduction

In the introduction of this chapter, we explain why we have chosen for a field study, in our quest for the 'real risk'. In section 5.2 we formulate research questions that represent steps on this quest, in 5.3 we give an overview of the three field studies we performed, in 5.5 we stress the anonymity of our respondents and in 5.6 we discuss the generalisability of our findings.

In section 1.1 we introduced the generic question of his thesis, whether the risk as assessed by the auditor on audit objects in his practice, represents the 'real risk' of a material error. For finding an answer to this question, two ways of action were considered:
1. an experimental approach, in which cases with a pre-specified risk would be administered to experimental subjects, and in which interesting variables could be investigated on their effect on the risk assessment by the subjects;
2. a field study (archival study), in which in principle on real audit objects two measurements would be performed: the level of risk as assessed by the auditor and one or more indicators of the 'real risk' associated with the annual accounts of the audit object. In this field study the source of the data would have to be the audit files of an audit case, because it is on these that an auditor bases his opinion and in which the exact outcomes of the audit process are given.

There were several reasons for choosing the second option.
The first, main and decisive reason was that in our opinion it is almost impossible to construct the "right cases". Such a "right case" should:

- entail a pre-specified level of risk;
- not give cues in the description of these cases as to the level of risk aimed at by the researcher;
- entail certainty as to the effects on the error rate (and as a consequence on the audit risk) of the included risk factors.

For to realise the third bullet for instance, it would be necessary to know of the effect of each risk factor. But the effect of such a factor is dependent on many other factors in a way that almost absolutely inhibits the construction of a "natural case". In other words, we did not see how to include cues pertaining to the 'real risk' in an audit object, at least satisfying the following properties:

- like in practice, they may not be that overt and they may not be that sure to have a bearing on the generation and/ or detection of errors;
- like in practice, they sometimes only can be observed by being present in the organisation; in other words, messy desks, chaotic archives, wavering communication can be described, but could be interpreted differently on direct observations;

- like in practice, they sometimes only can be discovered by talking to people responsible for the processes; the considerations of the previous bullet apply: a phenomenon being described or being directly observed may make a difference in interpretation.

In other words, we did not see how to include risk cues, without giving away too much information on their bearing on the audit risk in the case description. And we also did not see how to provide for the many ways in which risk factors can interact in creating and/ or hiding an error. Our discussion in chapter 3 on the complexity of risk assessment gave a more elaborate analysis of this complexity. This complexity also implies that it is very hard to judge whether a risk assessment by an experimental subject as based on the cues in the experimental case, be it low or high, is correct. Both low and high might be justified. For instance because experimental subjects may give a varying interpretation to that cue, due to their varying experience in similar cases. This may lead to different assessments which still are consistent with the given description.

A second and also decisive reason was that an experiment would not completely solve the question whether auditors in practice are able to give the right assessments of the audit risk, in particular the occurrence risk (OR). Obviously this second reason is related to the first, but it also stands on its own, for instance because in an experiment only a limited number of factors can be varied.

A third reason to choose a field study was given in 3.2.5: there it was argued that as far as heuristics and biases are included a purely experimental study would not answer the question whether the auditor can assess the ‘real risk’.

A fourth reason is given by Van Kuijck (1999) who aims at confirming findings from experimental studies in a quasi-experimental approach. He refers to authors who doubt the value of experimental studies because they have the suspicion that subjects act differently in experimental situations as compared to real-life situations. This suspicion is fully justified, since it refers to the equivalent of the well-known Hawthorne effect, which was discovered in 1928 (see Babbie, 1994 p. 236; Cook and Campbell, 1979, pp 39,60,66 and others).

All these considerations result in the conclusion that the generic question given in section 1.1 can not be answered by an experimental approach. So we chose an (archival) field study approach, which is consistent with the framework of Bonner (1999, p. 389). Bonner also mentions external validity as an advantage of archival studies, but with her, we realize that, by not adopting an experimental approach, causality is harder to find.

Next to the counter indications for an experimental study, two positive reasons for a field study can be stated.

1. The first reason is that a field study can answer the basic question "does it work?" without having to justify the experimental conditions. It measures "real life" by directly measuring the auditors risk assessment and the corresponding errors. Every property, seen or unseen, is included in the object on which the auditor has to assess the risk. Moreover, when audit files are the basis for the measurement of the auditor’s assessment, there is no measurement error, because the assessments filed are also the ones that were the basis for the audit design, in which the audit risk aimed at is included.
2. A second reason is that a field study still offers the opportunity to look for factors that influence the quality of risk assessment, like properties of the auditor,

properties of the audit object, properties of the audit environment. In the next section we will go into further detail.

5.2 Research questions

In section 1.1 we stated the following generic question:

"Does risk assessment by an auditor represent the 'real risk' of a material error?".

Due to the widespread use of risk analysis, we designed this study with the basic expectation in mind that risk assessment by auditors will certainly have some predictive value for the error rate (and its possible materiality) in the annual accounts. With this prior attitude towards risk assessment, we developed the key question in four directions:

1. the first direction regarded the relation between risk assessment and the four criteria developed in chapter 2;
2. the second direction regarded an exploration of possible moderator variables: variables that influence the strength of the relationship between risk assessment and the criterion variables;
3. the third direction was an attempt to improve risk assessment by decomposing the overall assessment into assessments of risk indicators;
4. the fourth direction regarded the predictive qualities of system tests[14] for the error rate in account.

We will present the resulting research questions in 5.2.2 through 5.2.5. But first, we will discuss the level of analysis that applies to our study.

5.2.1 Two levels of analysis: the organisation and the pooled organisations

Clearly the research questions regard the capability of the individual auditor in risk assessment. However, most of our data are derived from only one assessment per individual auditor. Therefore we had to pool our data to the level of the organisation (the auditor belongs to) and analyse them per organisation. This means that our pronouncements rather apply to the organisation than to the individual auditor. But as the organisation is made up by individual auditors we could say as well: "they apply to the auditors of the organisation". Moreover, it refers to the necessity for individual auditors in the organisation to assess their risks in such a way that the organisations uniform guidelines on the extent of substantive testing, given a risk assessment, result in an audit of sufficient size. It can be stated that the validity we will test at the organisational level is necessary for the use of such uniform tables and thus gives a sufficient reason for taking the organisational level as level of analysis.

These considerations lead to the choice of the organisational level as the most natural level for pooling. The same considerations also lead to the conclusion that one assessment per responding auditor is sufficient.

Where relevant we will also pool the data at the level of pooled organisations. Pooling makes sense, because many organisations base their audits on the same handbook, all auditors are subject to the same international auditing guidelines and standards, and the key question: "Does the auditor assess the 'real' occurrence risk?" is the same for all auditors/ organisations, regardless the underlying methodology. The advantage of

[14] "System tests" in this thesis is synonymous to "tests of controls", or "compliance tests". The term "system tests" is widely in use in the Netherlands and the research reported in chapter 9 is based on the Dutch practice, which makes the use of the term "system tests" natural.

such a higher pooling level is that our sample size increases and therefore the power of the statistical tests or the possibility to include (more) predictors in regression analyses. But the relevance is limited by the possibility that there is much variability between organisations: it could result in significant correlations for the pooled organisations (validity in general with respect to some criterion), but insignificant correlations per organisation (so no validity with respect to the same criterion for an organisation). Where the interesting question with regard to validity is at the organisation level, we have to be careful with the results for the pooled organisations. Moreover searching for causes is complicated with varying organisations and associated cultures and experience.

5.2.2 The relation with the criterion variables.

As argued in 2.3.2, a valid risk assessment may be expected to correlate positively with the position of the error rate in the audited account relative to the level of materiality. The first research question directly derives from this expectation:

Research question 1: To which degree will risk assessment correlate with the position of the error in a sample relative to the materiality?

As argued in 2.3.1, a valid risk assessment may be expected to correlate positively with the error rate itself in the audited account. The second research question directly derives from this expectation:

Research question 2a: To which degree will risk assessment correlate with the error rate in the audited account?
Research question 2b: To which degree does this correlation vary over organisations?

As argued in 2.3.3, a valid risk assessment may be expected to correlate positively with the sampling risk (SR) associated with the outcome in the audit sample, except for the possible influence towards a negative value for the correlation between SR and OR due to the dependency of sample design and OR, coming from the lower sample sizes and therefore higher SR, when OR gets a lower assessment. This dependency will lead to a positive correlation between OR and sample size.

In preview of the analyses in chapter 6 we observe that the correlation between OR and sample size is only .101, which mitigates the dependency of the correlation on the sample design. We also recall that the standardised sampling risk, based on the same sample size in all cases, does not suffer from this problem. So we can derive the third research question directly from the expectation of a positive correlation:

Research question 3: To which degree will risk assessment correlate with the sampling risk (SR)?

As argued in 2.3.4, a valid risk assessment will show varying distributions of the error rate for different assessed levels of risk. The fourth research question directly derives from this expectation:

Research question 4: To which degree will the distribution of the error rate vary with the level of assessed risk?

As explained in 2.3.1, the occurrence risk not only depends on the error rate, but also on the level of materiality: the higher the materiality, the smaller the occurrence risk, given an observed error rate. For groups of accounts with the same level of materiality, the relation between error rate and the occurrence risk only depends on the error rate.

In such groups and with valid risk assessment, the correlation of error rate and risk assessment will increase, compared to that in the whole set of accounts. This leads to the fifth research question:

Research question 5: To which degree will the correlation between error rate and risk assessment increase when calculated for groups of accounts with the same level of materiality compared to the correlation for the whole group of accounts?

In chapter 2 (2.3.3) we argued that in principle the (possibly standardised) sampling risk SR, given the error rate found, may be the best validation criterion for the occurrence risk (OR). For valid risk assessment it therefore can be expected that the relation between OR and SR will be stronger than that between OR and the 'audit position' or between OR and the error rate. This leads to the sixth research question:

Research question 6: To which degree will the occurrence risk (OR) show a relation with the sampling risk, which is stronger than that between OR and the 'audit position', or OR and the error rate?

5.2.3 Moderator variables

In chapter 3 various variables that improve or might improve risk assessment were discussed, among others: experience as an auditor and knowledge of the type of business the audit object belongs to. Literature (among others: Abdolmohammadi & Wright, 1987, Colbert, 1989, Stone & Dilla, 1994, Davis et al, 2000), showed that better assessments may be expected from a more experienced auditor and from one having more knowledge of the type of business. In the same references, implicitly complexity of the audit object [15], was shown to negatively influence the quality of risk assessment. For more complex audit objects, less valid assessments may be expected.
These references and associated expectations imply that the corresponding variables are potential moderator variables for the relations that are of interest to us.

We did not take size as a possible moderator, because we found no literature to have a basis for its moderating effect. Moreover results on its (possible) moderating effect are hard to interpret: size of an account (the total number of monetary units), may have both a positive and a negative influence on the quality of risk assessment. Positive because with a larger size errors of a certain percentage get larger and therefore more care is taken not to miss them (which is also mirrored in a level of materiality of a lower percentage). The influence may be negative, because larger may mean more complexity, making risk assessment more difficult. It also may be indifferent for size, because it is not the size in MU's that determines the difficulty of performing a risk analysis, but the 'size' (say in number of and relations between controls) and related complexity of the system of controls. So, whatever the outcome of an analysis, there is always a justification for it; 'size' as a moderator is not informative, if no other variables can be included in the analysis.

We introduce the following research questions with respect to moderator variables.

With growing complexity organizations are harder to evaluate in terms of assessed risk; the auditor will be aware of that and put the effort for the assessment at a relatively high level, in order to overcome the complexities. This might compensate for the complexity,

[15] Variables like 'complexity of the audit object' and 'experience' have to be defined and operationalised to make them measurable; we will do this in chapter 6, when we report the findings as to these research questions.

were it not that human mind has great difficulties in combining assessments (see for instance Burgstahler et al, 2000). So in the end the complexity may be expected to reduce the level of validity. This leads to the 7th research question:

Research question 7: To which degree will the validity of risk assessment decrease with the complexity of the audit object?

With more effort an assessment in general may be expected to improve. This leads to the 8th research question:

Research question 8: To which degree will the level of validation increase with the effort put in the assessment of OR?

As a consequence of Abdolmohammadi & Wright (1987), Bell & Carcello (2000), Davis et al (2000), Stone & Dilla (1994), Burgstahler et al (2000), Smith & Kida (1991), Joyce & Biddle (1981a), Johnson et al (1991), (see chapter 3 (3.4.2.2 and 3.4.3)), it is interesting to look at experience as potential moderator variable. In many cases we know whether the audit regarded a first or a repeated engagement. This stands for less or more experience with the business and to a lesser extent for experience as an auditor. Because our data do not distinguish between these two types of experience, the differential influence (indicated by some of the studies in chapter 3) of experience in the business and as an auditor, can not be confirmed by our analysis . Still it is worthwhile to add research question 9:

Research question 9: To which degree will the level of validation increase with the experience of the auditor with the business?

The following considerations regard a potential moderator variable that is only nominal by nature. Still it is interesting to look whether it has influence, even if the direction of the influence cannot be predicted.

An auditor attains a certain level of experience in risk assessment in interaction with his or her direct colleagues. The firm's or department's manual will formulate an assessment strategy which in its turn, in conjunction with the exchange with colleagues in the organization, will cause assessments by varying auditors within organizations to be interchangeable and therefore to be at a relatively stable level of validity. The question is whether this level is also stable over organizations. There are reasons to believe that this will only be the case to a more modest extent; we mention the following:

- there is lack of professional exchange,
- audit manuals are secret outside companies,
- the audit manual plays a central role in auditing,
- over- or under-assessments are not made public.

These (and possibly even more) reasons will not cause risk assessment to be totally different over organisations, because all organizations are subject to international guidelines on auditing and auditors in general share education; moreover there is some exchange for instance by way of published research. These considerations lead to the 10th research question:

Research question 10: To which degree will the level of validity vary over organizations?

The possibility that there is also variation in validity between individual auditors within one organization cannot be tested in this study, because only one audit per respondent was collected.

5.2.4 Decomposition of risk assessment

In chapter 3 the complexity of the assessment task was stressed. In more complex situations, heuristics like anchoring, availability and representativeness are more likely to lead to corresponding biases. Several authors (Colbert, 1989, Bell & Carcello, 2000, Buckless, 1989, Abdolmohammadi & Wright, 1987) actually showed that complexity has influence, and also that it sometimes can be dealt with, a) by making assessments at the assertion level (for instance Waller, 1993, Buckless, 1989), b) by excluding judgment as far as possible: logistic regression on directly measurable risk factors performs better than judgment of auditors (Bell & Carcello, 2000), c) by decomposing the assessment needed into part- assessments (Roberts & Wedemeyer, 1988, Wallace & Kreutzfeldt, 1993).

In this study the possibility to improve risk assessment by decomposition of the assessment task also has been investigated. The idea was to decompose the assessment task into a set of subtasks, each of which would be less complicated. Each subtask consisted of giving an assessment on a 'risk indicator'.

Definition of risk indicator
A risk indicator is an aspect of the audit object or its context, that is expected to affect the risk that the administrative process will produce annual accounts with a material error.

In appendix 1 (the questionnaire) we explain how we came to the set of indicators, that was used in the questionnaire.

Indicators imply judgment.
The respondent was asked to give his judgment with respect to each indicator as to the question whether the audit object would have less or more risk of a material error or would be risk-neutral. This question often will produce the same answer as the question of the quality with respect to that indicator. But by including judgment in the indicator with regard to the effect on the risk of a material error, an assessment is explicitly asked. This allows for judgments of the auditor that, although the quality may be poor, the resulting risk still may be low in the specific situation. It means that one of the central properties of judgment is included in the way we use risk indicators: it is more than just looking at the parts and it allows intuition or professional judgment.

This approach also is consistent with a goal of our study: to look whether assessment improves when the situation gets less complicated (section 3.6, 3rd question). For that we have to measure risk assessment at two levels: at the level of the usual audit risk model next to the level of the risk indicators.

Decomposition of risk assessment asks for the aggregation of the assessments per 0indicator. In principle we chose for aggregation by way of regression analysis. By using this technique, it would be possible to find which indicators are most important in risk assessment. Also variation in these models over organisations could be found. And, not the least, the explaining power of the risk indicators can be compared to that of classical risk assessment. We chose for the error rate as the dependent variable, because it appeared that classical risk assessment correlated strongest with this validation criterion.

Construct validity of the risk indicators
In the course of the study the opportunity was taken to involve the participating auditors in controlling the construct validity of the risk indicators. When we introduced our study to a participating organisation, we discussed the questionnaire, in particular the risk indicators, with our contact persons. In this discussion the view of the auditors on five desirable qualities (mentioned in appendix 1) was especially asked for. As a rule, only minor changes resulted from these discussions; in one case one indicator was added, in another one was deleted. In the end this did not play a role in our findings, because we used a set that applied to all organisations, as will be made clear in chapter 7.

Next to the analysis of correlations between and of regression of the risk indicators on the occurrence risk, this check on consistency can be seen as a triangulation on the construct validity of the risk indicators (see Babbie, 1995).

Research questions concerning risk indicators
The above can be summarised in six research questions:
The first three derive from the question whether the risk indicators are consistent with the classical audit risk model.

Research question 11: Are the bivariate relations between risk indicators and the occurrence risk in the expected direction and of sufficient strength?

Research question 12: Can the occurrence risk be predicted from the risk indicators by way of a regression model?

Research question 13: Can the risk indicators be seen as an appropriate representation of the view auditors have on audit risk assessment?

The next three derive from the question whether the predictability of the error rate can be improved by the use of risk indicators. This question can be decomposed as follows.

Research question 14: Are the bivariate relations between the risk indicators and the error rate in the expected direction and of sufficient strength?

Research question 15: Can the error rate be predicted from the risk indicators by way of a regression model?

Research question 16: Is the explaining power of the risk indicators larger than that of the classical risk assessment?

5.2.5 System tests as predictors of errors

In section 2.4 we argued that the use of system tests to underpin risk assessment when it is used as a replacement of direct substantive testing, implies a question as to the validity of system tests. We also mentioned that we would test that necessary condition for system tests to be valid as underpinning of risk assessment.

System testing serves to establish the proper operation of the system of controls; risk analysis serves to assess the quality of the system of controls, combined with assessment of the influence of the context in which this system operates on the risk of a material error, in order to assess the occurrence risk. Related to the difference in goal, there are three essential differences between risk assessment and system testing: (a) the level of assessment, (b) the extent to which judgment plays a role and (c) the extent to which the design and the operation play a role.

a. The level of assessment
Risk assessment has the annual accounts (or a sub-account) as its object and assesses the risk (of a material error) at the level of this (sub)account. When investigating its validity, only a measure at the level of the (sub)account is relevant. System tests, by definition, have the individual transactions as their object. So investigating their validity should be done at the level of a transaction. The most logical criterion for validity of a system test then will be the error in the transaction, as stated in section 2.4. This means that we need *dual purpose tests* to validate system tests. A dual purpose test is a test on an individual transaction in which the operation of the controls (the administrative system) is checked (also called compliance testing, see e.g. Arens & Loebbecke, 1997) and in which also the correctness of the book-value (the monetary value for which the transaction has been booked) is investigated (also called substantive testing, see idem). The logical expectation in such a dual purpose test is: the better the system has operated, the more probable a zero error or a small sized error.

b. The role of judgment
In risk assessment, judgment plays a prominent role: the auditor forms an image of the administrative processes, analyses them on weaknesses, looks for the controls that should neutralise these weaknesses, wonders how well they are designed to meet them, gets an impression of the qualities of the personnel, tries to get an impression of how well the procedures are followed etc etc. In system testing, judgment is far less important and observation is the key activity: it is observed, how well the procedures are maintained in an individual transaction. This observation is done on cues, directly related to the procedures, such as the signatures of employees that correspond to separation of duties.

c. The extent to which the design and operation play a role.
In risk assessment both the design and the operation of the administrative procedures are assessed, with emphasis on the design, as outlined in the previous point. The operation is also paid attention to, but only in so far as the "existence" (as is said in the Dutch auditing practice, see: Leerboek, 2003) is concerned. “Existence” is established by ‘walk-through’ tests: with one or a few transactions every relevant part of the administrative procedures is tested. When this test is positive, the administrative organisation "exists": it has been shown that the designed controls "exist" and can do their work (see for instance: Leerboek 2003, Arens & Loebbecke, 1997, ISA 400 par.15). But establishing that the controls continually operate in a proper way, asks for a lot more system tests; system testing is meant to meet this necessity.

So in system testing the controls are tested on their correct and continual *operation*, *given their design*; this is done by testing the controls in a sample of transactions of appropriate size. In principle, when poorly designed administrative procedures operate well, this could lead to excellent outcomes of the system testing. So system testing on its own is not a sufficient indication for the quality of the administrative procedures. But note that when an auditor establishes the controls to be poorly designed, he will not choose to do extensive system testing, because that does not add to the assurance regarding the absence of a material error.

These differences between system testing and risk assessment leave unaffected that they both aim at assessing the quality of the administrative processes that produce the annual accounts. In this ‘one-two’ system testing plays the role of underpinning the risk assessment, by showing the strength of the operation. We could also say that it completes risk analysis, by showing whether the procedures operate properly, given their design. When system testing confirms the assessment of "low risk", the auditor

may decrease his substantive testing; consequently substantive tests are replaced by system tests. It is a widespread practice to do so. We may conclude that this decrease asks for *two* necessary conditions to be met: proper design and proper operation of the controls.

Replacing substantive tests by system tests implies that, just as to risk assessment in general, also to system testing qualities are attributed as to having predictive power for the error rate. So if system testing would appear not to be a predictor of the error rate, risk assessment cannot be expected to work. It follows that the predictive qualities of system testing are worth to be investigated. The 17th research question applies to this.

Research question 17: To what extent are system tests predictive for the error rate and valid in that sense?

For the study into this validity of system testing, we again chose the option of a field study, because we could be provided with the results of many thousands of dual purpose tests. See section 5.3

5.3 Three field studies

Our research consisted of three field studies.

Five audit departments of Dutch ministries, two private audit firms and one audit office of a European country participated in the **first study** (see chapters 6 and 7). They were asked to give information on audits from 1996 or 1997, by way of filling out a questionnaire (see appendix 1), which contained questions for an audit object related to:

- global factors that determine the initial risk assessment
- risk assessment in the classical sense, IR and ICR explicitly combined into the occurrence risk OR;
- assessments on 23 risk indicators;
- the errors found in the audit object;
- some background variables (type, size in monetary units, etc.), which might be moderator variables.

It was explicitly asked to answer the questions with the audit files as a basis. The auditor who did the audit was asked to give the information, all coordinated by our contact person in the organisation.
At the public audit organisations the audit objects regarded either a part of the direct expenses of the related ministries or of the expenses of some agency. At the private organisations mostly accounts receivable or stocks were the audit object.

In the **second study** (see chapter 8) four audit departments of Dutch ministries participated. The aim of the study was to check the findings in the first as to the unstable relation between occurrence risk and error rate. We collected the data ourselves, by examining the audit files, concerning 2001, in these audit departments for:

- risk assessment in the classical sense: IR and ICR;
- the error rate;
- background variables.

When these data were not in the files, some effort was made to retrieve them, but none to reconstruct them.

The **third study** (see chapter 9) was only possible because we were provided by a Dutch ministry with more than 30,000 records containing data on the dual purpose tests over five years (1995-1999). The aim of this study was to check the predictive qualities of a system test for the error rate in the corresponding transaction.

The first study was optimistic about the relation between error rate and risk assessment, especially in the possibilities of decomposition of risk assessment into risk indicators. But among many other things, it found an unstable relation between risk assessment and error rate. The second was meant to replicate the first study especially with an eye on the unstable relation between classical risk assessment and error rate (and hopefully find more stable results). The third was meant to find a way out of the unsatisfactory findings of the first two studies, by trying to find a confirmation of the relation between system tests and error rate. If this relation would be found, this would mean that at least the part of risk assessment in which system tests underpin risk assessment, would be valid.

Only the first study was planned at the start of this project, the second and third were entered into on the basis of the findings of the first and second study.

5.4 Anonymity

An arrangement was made with the cooperating organisations, where we promised to process their information and publish the results in anonymity.

5.5 Generalisability

The key question of this project is "does risk assessment work; does it indicate the real risk". This question can be asked at various levels:

- at the level of the auditor;
- at the level of an auditors' organisation;
- at the level of organisations in general.

Relevant questions as to this working are:

- does it depend on type of clients' organisation;
- does it depend on some moderator variables.

Findings in the studies reported in his thesis do not apply to the level of the auditor: for that purpose it would have been necessary to have repeated measurements of risk assessments per auditor.

Findings in the studies do apply to each separate auditors' organisation. In principle the risk assessments that were the basis for this study may be supposed to be representative for risk assessments in general by auditors of the organization, although no random procedure for selection of auditors and/or audit objects was used. But there was no selection of the respondents on their (alleged) risk assessment capabilities. Decisive for their cooperation was whether they dealt with audit objects aimed at and that, especially with the private firms, they were willing to spend their costly time on it. But there are restrictions as to this applicability: the sample of audit cases from each organisation was selected by the organisation itself and there was a criterion for selection that only accounts in which some error was to be expected were wanted. As a matter of fact there are good reasons to assume that the organisation did not select cases as to get a positive relation between risk assessment and error rate, because the

organizations did not keep a systematic record of this relation. Moreover, they would have had a difficult job to select with that purpose, especially with respect to the risk indicators, because there are so many risk indicators on which the relationship is dependent. So the findings will be generalisable to the same type of audit object. The findings will also provide for a significant indication of the existence of the relations studied: if they do exist, it is hardly imaginable that we will not find them in a sample of so many cases. It is also plausible that the findings will apply to some extent to other types of sub-accounts within the same organisation. When, for instance, a class of accounts receivable is assessed at a low risk and still a material error is found, this increases the probability that the same might occur in another type of sub-account. Apparently in such a case, the quality of risk assessment in this audit organisation allows such an unexpected match. Note that also a correct assessment 'low risk' does not exclude the existence of a material error.

Findings in the study will also apply to other auditors' organisations, because the participants in this research were not selected with an eye on their risk assessment qualities (had they been known). There are also restrictions to this applicability, as the sample of organisations neither is random. But it can be argued that the participating organisations act under the same audit rules and regulations, guidelines, international standards, professional education as the ones that did not participate. These conditions will tend to cause a certain degree of uniformity in risk assessment over organisations. If these conditions would not work at all, not only would the generalisability of our findings to a larger set of organisations be problematic, but also the applicability and working of regulations, guidelines and standards. And moreover, the psychological mechanisms that are discussed in chapter 3 equally apply to every human being.

As far as generalisation is aimed for, we could say that there is an implicit population to which the results apply: that of similar accounts, auditors and organisations, and that it is highly improbable that this population is empty. But in fact in this project generalisability was not to the main goal. The way of reasoning was rather the other way round: if in general risk assessment works, it will also work in the participating organisations, especially if they are not selected for their qualities in assessment. So if it works, some signs of it must be found. Reversing this argument means that when no relevant signs are found it only can be concluded that there is something wrong with risk assessment.

We can also state this as follows: the generalisability is to the validity of the audit risk model as such, and not to the whole population of organisations using it (see for instance: Yin, 2003, p.10). But in such a statement we still have to be aware that 'generalisation to a theory' as Yin calls it, only makes sense when a population exists in which the theory really can be shown to work. So our 'existence of an implicit population' and Yin's 'generalisation to a theory' may be equivalent.

Chapter 6: Classical Risk Assessment at eight Institutions.

Four relatively complete actions in three studies were taken in our research into the 'real risk': (1) validation of the classical assessment of OR, (2) an attempt to improve classical assessment of OR by decomposition of the assessment over risk indicators (1st and 2nd action formed the first study), (3) a replication of the validation of classical assessment of OR (in the second study) and (4) investigation of the predictive power of system tests for the error rate (in the third study). Varying validity of classical risk assessment per organisation (1st action) and problematic results with risk indicators (2nd action) led to the choice for the replication (3rd action).
In this chapter we report the results of the 1st action.

6.1 Introduction

The leading question of this chapter is how well an auditor's assessment of the occurrence risk (OR) will perform at each of the four validation criteria used in the study: In the sections 6.2 through 6.5 we will investigate this quality, firstly for the pooled organisations and secondly at the level of distinct organisations. By investigating possible moderator variables in 6.6, we will answer the remaining research questions of the eleven that are dealt with in this chapter. We will give a discussion of the findings for every single research question and end with an overall conclusion and discussion in section 6.7 and a summary of our findings in section 6.8.

The four validation criteria were introduced in chapter 2 where, in section 2.3, we chose:

- the "audit position", the position of the error rate relative to the materiality,
- the error rate, as estimated by the auditor
- the sampling risk
- the conditional distribution of error rates.

All criteria are related to the error rate, which relationship we can recapitulate in the following table

Table 6.1: The four validation criteria

Validation criterion →	'audit position'	Error rate	Sampling risk	Distribution
Relation to error rate(s) →	Is error rate larger than materiality?	Error rate itself	What is the risk that the real error rate exceeds materiality, given the error rate found?	What is the distribution of the error rates for distinct levels of OR?

The findings regard our research in the first study. In this study we collected the data concerning annual accounts from 1996 or 1997, by way of a questionnaire in which the respondent was asked to give the values of the assessed risks, the error rates found and other variables, by retrieving them from the audit files. With every organisation, the questionnaire was discussed in advance, with the contact person and some of the (other) respondents. The data regarded among other things:

- the occurrence risk (question 3.3), for which he probably had to transform the assessment of IR and ICR (in the audit files) into OR.

- the error (q 4.3); because of the various possible forms in which the error could have been filed, we gave the respondent various ways to give the error: as an amount or as the sum of taintings; in our data processing we calculated the estimated error rate from these data dependent on the sample design as: the average tainting (in case of monetary unit sampling), or the estimated amount of error divided by the size of the financial statement (in case of line item sampling), or simply the amount of error found, without extrapolating it, also divided by the size of the financial statement; (in case of judgmental sampling). When the sampling was judgmental, we always checked how it actually had been performed, in order to apply the right estimation procedure,
- the size of the financial statement (in monetary units) (q 1.2),
- the materiality (q 2), which could be given as an amount or as a percentage of the size (or both),
- the size of the sample and the number of line items in error (q 4.3).
- the potential moderator variables as introduced in 5.2.3 (effort, experience,(q. 3.3),, complexity, (one of the risk indicators))
- the risk indicators (q 3.1).

(For the questionnaire, see appendix 1).

In all, as mentioned in chapter 5, eight organizations participated in this first study of our research: a national court of audit from one of the EU-countries, five audit departments of Dutch ministries and two private audit firms, also in the Netherlands. With each organisation the questionnaire was discussed with a contact person and with participating auditors. One of the leading questions was whether the questionnaire represented the way an audit was done by the organisation. Sometimes this led to minor adaptations of the questionnaire, but not to an extent that it changed the substance. The questionnaires were distributed and collected by our contact person in the organisation. The organisations were explicitly asked to fill out the questionnaire with data from or based on the actual audit files.

On average, about 20 cases per organisation were collected. Data of the three organisations with the smallest number of cases together, only consisted of 13 cases. Because in these organisations the same handbook (that of governmental audit organisations) was in use and because the audit objects of these organisations were very similar (in numbers of transactions, in size and kind of transactions, subject to relatively similar regulations) we pooled the data into one unit of analysis which we will be referred to as ‘the triplet’. For the analysis this triplet was treated as a distinct organisation. So 8 organisations formed 6 units of analysis: 5 organisations and one “triplet”. Whenever there were questions concerning data, direct contact between researcher and respondent solved these.

The data regarded the annual accounts of 1996 or 1997

All data were analysed at the level of the (6) units of analysis. We will refer to this level as to the “level of the distinct organisations”, thereby disregarding that one of these ‘organisations’ is a triplet. The data of all eight organisations were also pooled and subjected to similar analyses as at the level of the organisations. So actually, the pooled organizations formed a seventh unit of analysis. We refer to 5.2.1 for a justification of these levels of analysis.

6.2 Risk assessment and 'audit position'

6.2.1 Definition of 'audit position'

In 2.3.2, we defined ''audit position'' as follows:

Definition of 'audit position':
The 'audit position' of the error rate is its position relative to the materiality; in our study it is labelled 'not OK', if it is larger than materiality; it is labelled 'OK' if the error rate is smaller than materiality.

In our analysis, we defined a corresponding variable 'audit position' which takes the value 0 (not OK) when the MLE (most likely error) is larger than materiality and the value 1 (OK) when the MLE is smaller than materiality. We refer to this latter position as 'unqualified', because of the clear analogy to the audit opinion in case of the complete annual accounts.

We recall our remark at the end of 2.3.2 where we stated that the 'audit position' may be OK ('unqualified') because the most likely error (MLE) is smaller than materiality, where in the same case the audit opinion would have to be qualified, because the upper error limit (UEL) is larger than materiality. In fact, if the MLE would be taken as sufficient for an unqualified opinion the reliability of that opinion could be only approximately 50%. This, because roughly the most likely error has 50% chance to be smaller than the real error, if in fact the real error just exceeds materiality. So when a most likely error just below the level of materiality would actually lead to an unqualified opinion, this only would have this same reliability of approximately 50% (instead of the standard 95%, aimed for in the audit practice). If we would define 'audit position = OK' with a rule that implies a larger reliability, more 'not OK positions' would result (in the same situations), because then only most likely errors which are sufficiently smaller than the level of materiality, would result in an 'OK position'.

In our third validation criterion: the sampling risk, this dependency on materiality and the corresponding reliability is accounted for.

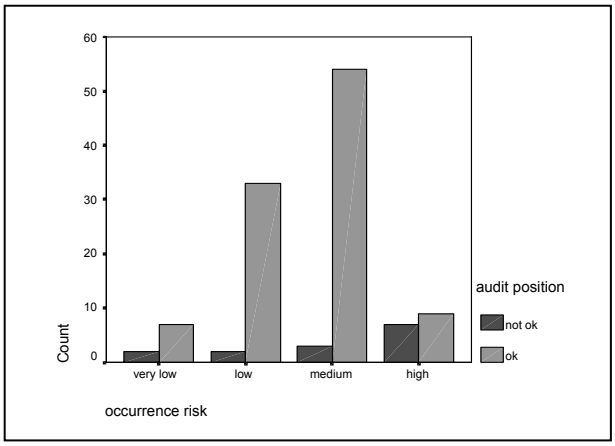
6.2.2 Results for the pooled organisations

In case of a valid risk assessment, it may be expected that the relative number of cases in which 'audit position = OK' decreases with increasing OR. The following cross tabulation shows the quality of the assessment of OR in this respect. It is illustrated with a bar diagram in figure 6.1. As can be seen, two (of nine) 'not OK positions' are found at OR=very low and also two at OR=low. Moreover, the rate of 'not OK positions' decreases up to OR= medium, where an increase should be expected (see the row percentages). Only at OR=high the picture is more in line with the logical expectation. The chi square statistic had a very significant ($p<.005$) value of 20; but this does not imply the expected relation. This, because the chi square statistic only accounts for deviations from the expected cell frequencies; these deviations can be such that they point in the expected direction in one row but in the opposite direction in the other. For chi square the deviations count, not their direction.

Table 6.2: Occurrence risk by 'audit position' for the pooled organisations

'audit position'→ Occurrence risk↓	Not OK	OK	Row totals
Very low	2	7	9
	22.2%	77.8%	100%
Low	2	33	35
	5.7%	94.3%	100%
Medium	3	54	57
	5.3%	94.7%	100%
High	7	9	16
	43.8%	56.3%	100%
Column totals	14	103	117
	12.0%	88.0%	100%

Figure 6.1: 'audit position' by occurrence risk for the pooled organisations



The cross tabulation shows a relation that certainly is not strong, but the relatively many 'not-OK' for OR='high' indicate the possibility of a (weak) relation in the expected direction. We further investigate this by calculating two correlation coefficients: the Kendall rank correlation and the point-biserial (PB) correlation between OR and 'audit position'. We should expect a negative correlation: with a change from 'audit position'=0 ('not OK') to 'audit position'=1 ('OK') the occurrence risk should change from relatively high to relatively low. The result is shown in table 6.3, last line. It turns out that both the K-correlation and the PB-correlation indicate a relatively weak relation in the expected direction; neither is significant at the 5% level.

6.2.3 Results for the distinct organisations

We applied the same analyses to the distinct organisations. The following table (6.3) gives the results of the K- and the PB-correlation of the OR with the row%% for the 'audit position'.

Table 6.3: K- and PB-correlations 'audit position' x OR by organisation

Organisation	K-correlation	p-value[16]	PB-correlation	p-value	n
1	-.35	.27	-.37	.27	13
2	-.014	.97	.15	.50	22
4	.085	.70	.057	.81	20

[16] A 'p-value' of an outcome of a statistic is the probability that this statistic might be equal to or larger than this outcome, given some null-hypothesis. When the p-value is small, the outcome can or will not be attributed to chance.

Organisation	K-correlation	p-value[16]	PB-correlation	p-value	n
5	-.29	.29	-.29	.20	21
6	.082	.69	.097	.65	24
8	-.69	.00	-.69	.00	17
Pooled	-.18	.13	-.18	.06	117

The results shown in table 6.3 are not very satisfactory, with an exception for organisation 8. Only the correlations for this organisation are significant at the 5% level (p-value<5%). But taken over all organisations, only 6 out of 12 correlations (K- and PB correlations pooled) are negative, the expected direction, so there is no systematic direction of the correlation over the organisations. Moreover, none of the other organisations shows a significant correlation.

For organisation 8 a closer look at the data (table 6.4) reveals that the relation between OR and 'audit position', as found, is strong: all 5 'not OK- positions' occurred for the 8 cases with OR='high'; all 9 OR='medium'–cases had 'OK-positions'. This relation was significant at the .007 level (Fishers exact test for a cross tabulation) and fully consistent with the correlations in table 6.3.

Table 6.4: 'audit position' by OR for organization 8

'audit position'→ Occurrence risk↓	Not OK	OK	Row totals
Medium	0	9	9
high	5	3	8
Column totals	5	12	17

Conclusion research question 1: validity with respect to 'audit position'
Assessment of OR only shows insignificant correlation with the variable 'audit position' for the pooled organisations; only one organization shows a significant correlation.

Conclusion research question 10: varying validity with respect to 'audit position'
There is considerable variation in the correlation between 'audit position' and OR; it is significant for one organization (8).

Discussion
As can be seen from table 6.2 there are only 14 cases in total, in which 'audit position=not OK' . This means that there is only a limited variation in the 'audit position' and that therefore the correlation between the variables 'audit position' and OR is very sensitive to only small changes in the number of 'audit position=not OK' on some level of OR.

6.3 Risk assessment and error rate

In chapter 2, we concluded that the error rate in itself is an indicator for the level of risk associated with the administrative process that produced the account under audit. Therefore it makes sense to use the error rate as a validation criterion for an auditor's risk assessment. In principle a correlation coefficient is fit for investigating the relation between risk assessment and error rate. But there are two complications:

1. The variable 'OR' is of an ordinal level, the variable 'error rate' is of a ratio level; therefore the Pearson product moment correlation in principle is not fit and so the Kendall or another rank correlation should be used. We will use both because, especially with larger sample sizes, the Kendall correlation and the Pearson correlation tend to show p-values close to each other. This is a consequence of the 91% power-efficiency of the K(endall)-correlation (see Siegel &

Castellan,1988, p. 254). When both correlations show similar p-values, that of Pearson can be seen as valid, and it makes sense to use the associated explained variance as an extra indicator of the strength of the relationship.

2. The variable 'error rate' is not normally distributed and may show serious outliers. This may cause significance tests to be invalid. Therefore we will examine the scatterplot of error rate by occurrence risk, as to the existence of outliers. It will indicate outliers that have to be skipped from the analysis. When the K-correlation is the basis for inference, this problem does not apply.

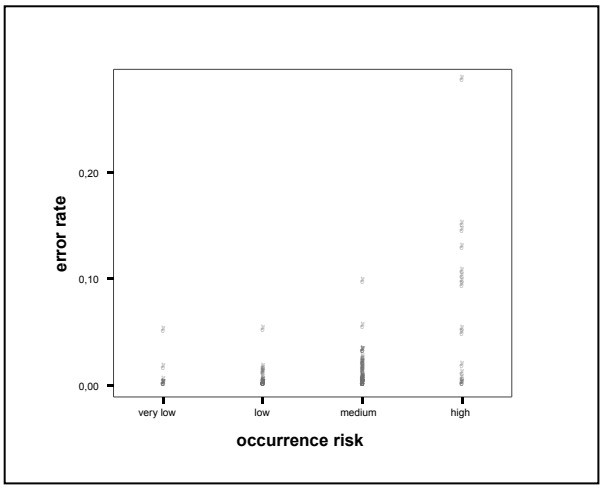
It should be noted that even these considerations regarding the data may fail to meet the probably pathological properties of auditing data. Willemsen (1996) shows that the estimated extreme value index (see Leadbetter et al, 1983) of real life auditing data comes very close to boundaries where even the central limit theorem no longer applies. In the very limited number of studies that investigate the statistical properties of auditing data (see for instance Tamura at all, 1989) these properties are also recognised, but not to the extent found by Willemsen.

The conclusion of this discussion must be that only the Kendall correlation leads to valid statistical tests, where the Pearson correlation only may help interpret the data when the values of both are similar.

6.3.1 Results for the pooled organisations.

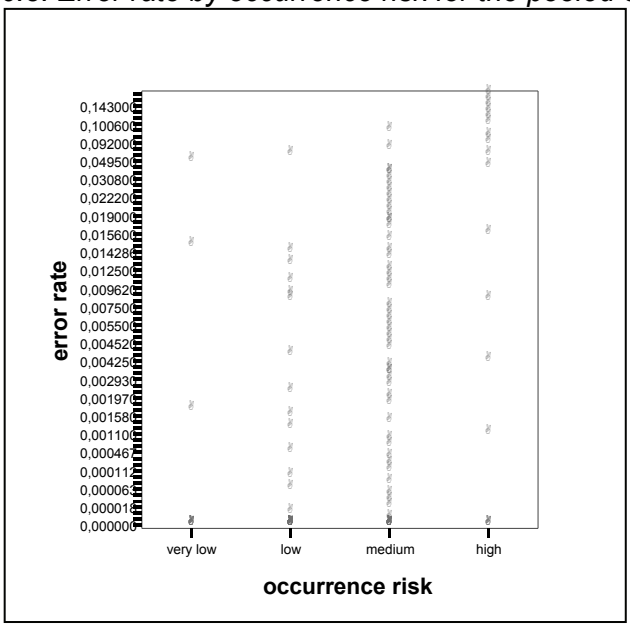
We start with a scatterplot of the error rate by the occurrence risk. Figure 6.2 shows the results.

Figure 6.2: Error rate by occurrence risk for the pooled organisations



The scatterplot shows one evident outlier in the right upper corner. It also shows a high density of error rates close to 0. This can be made more visible in a scatterplot with a scale for the error rate which is adapted to its actual distribution. This is shown in the next figure where the error rate is treated as an ordinal variable.

Figure 6.3: Error rate by occurrence risk for the pooled organisations



For the pooled data the Pearson and the Kendall correlation showed the following values:
P-correlation = .43 (p-value two-tailed: .00; n = 119)
K-correlation = .40 (p-value two-tailed: .00; n = 119)
Skipping the outliers mentioned above hardly changed the P-correlation. The correlation is satisfactory and highly significant, the corresponding explained variance of 18% (the square of .43) can be seen as moderate.

The fact that the occurrence risk OR only has four classes may inhibit a stronger correlation coefficient, because of the many ties this will cause. Therefore we also compared the means of the error rate per class of OR and calculated the association measure eta. This resulted in the following table.

Table 6.5: Means of error rate by OR for the pooled organisations.

OR	Mean of error rate	N	Std. Deviation
Very low	.0073	9	.017
Low	.0033	35	.0090
Medium	.010	58	.016
High	.072	17	.077
Total	.017	119	.039

The following association measures were calculated.

Table 6.6: Association measures error rate x OR for the pooled organisations

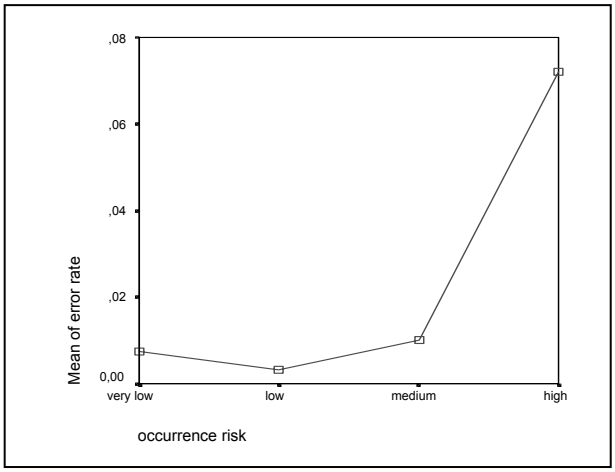
	R	R squared	Eta	Eta squared
Error rate by OR	.427**	.183	.590	.348

** significant at the .01 level

The means of the error rate in the 4 OR-classes turn out to be significantly different (p-value is 0) and the eta is .590; the explained variance -eta squared- is 35%. This is much better than the 18% variance explained by the P-correlation (R in table 6.6). But this improvement mainly is a consequence of the nonlinearity that eta allows.

Figure 6.4 shows the results graphically.

Figure 6.4: Means of error rate by occurrence risk for the pooled organisations



We see that with increasing OR, the mean of the error rate at first very slightly decreases and that from OR=medium to OR=high the differences increase drastically. We applied Tamhane's multiple comparisons test, which assumes unequal variances, to do pairwise comparisons of the means at the various levels of the occurrence risk. This resulted in table 6.7, which shows that the test shows one significant difference at the 5% level: that of OR = 3 and OR = 4.

Table 6.7: Pairwise differences for the means of the error rate, by OR.

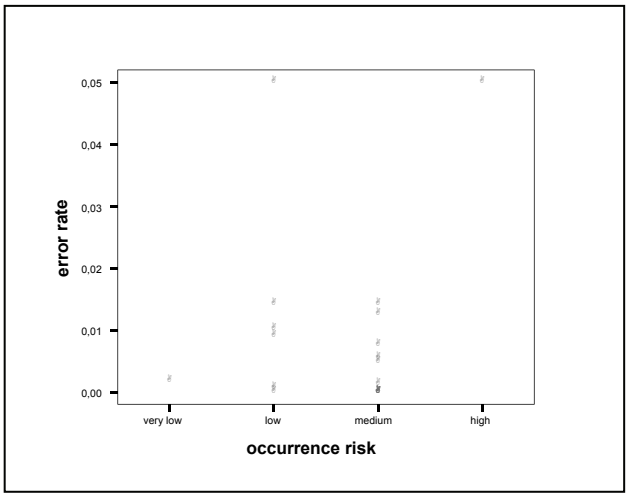
Difference for		Difference of means	Standard error	p-value
OR = 1	OR = 2	.0040	.012	.98
OR = 2	OR = 3	-.0068	.0068	.059
OR = 3	OR = 4	-.062	.0087	.027

We can conclude that the fact that eta is considerably larger than the P-correlation shows that the relation between the error rate and OR, lacks the linearity to be expected; it even lacks monotony. Where standard audit procedures allow less effort in substantive testing when OR is assessed at a lower value, this monotony is assumed. We could say that OR assessment is seen as a part of the assurance which is given in the audit opinion. This being so, implies that the lack of monotony must cause a problem: the audit opinion for OR='very low' might lack sufficient reliability. As a matter of fact it is hard to be precise on this possibility, because the tables and the audit effort they indicate are not calibrated. When we analyse the OR assessments in relation to the sampling risk, more can be said about it (see 6.4.2 discussion of ' ineffectiveness').

6.3.2 Results for the distinct organisations.

Just as we did for the pooled data, we plotted OR against the error rate and computed the correlations OR by error rate for the distinct organisations (with the three smallest combined into the triplet of size 13). All scatterplots showed more or less extreme outliers. We show the most interesting scatterplot, that of organization 4, which is associated with a moderately positive P-correlation and a moderately negative K-correlation. The results are given in figure 6.5 (for organisation 4) and in table 6.8 (for all organisations).

Figure 6.5: Scatterplot of error rate by OR for organization 4



The scatterplot shows how the data point at OR=high causes the inversion of the sign from the K- to the P-correlation (table 6.8). For the P-correlation this value will have a relatively large positive influence on the correlation coefficient. But for the rank correlation, its size is far less important; the relatively many small values of the error rate for OR = medium will result in the negative rank correlation.

Table 6.8: Correlations error rate x OR by organization

Organisation	P-correlation	K-correlation	n
1	.39	.30	13
2	.28	.22	22
4	.12	-.11	21
5	.52*	.55*	21
6	-.01	.16	25
8	.72**	.51*	17

** significant at the .01 level * significant at the .05 level

It can be seen that:

- almost all correlations are positive and two of the P-, together with two of the K-correlations are significant at the 5% level (one even at the 1% level);
- the differences between the P-correlation and the K-correlation are more substantial in some cases (see organisation 4, 6); this was already shown to be due to outliers, (figure 6.5);
- there is a substantial variation in the correlation over the organizations; most correlations are positive, but one P-correlation (organisation 6) and one K-correlation (organisation 4) are negative.

Intermediate conclusion research question 2: validity with respect to error rate
We can conclude that the overall picture shows a varying validity with respect to the error rate:

- the overall correlation (.427, table 6.6) is satisfactory
- four organisations show a positive correlation,
- only two (5 and 8) have a significant correlation,
- but organisations 4 and 6 show negative correlations.

So, even if this result would be interpreted as a moderate indication that the assessment of OR is valid with respect to the error rate, the substantial variability over organisations weakens this overall picture. Because, the variation over organisations implies that an auditor of a random organisation cannot be sure that in his organisation

risk assessment is valid. He will have to check this empirically. In this checking he will also have to investigate the monotony of the relation of OR and error rate.
We must conclude that:

An organisation, or department, or branch will have to assess independently of others whether this validity of its risk assessment is sufficient.

6.3.3 Controlling for materiality

So far we did not take the fact into consideration that the error rate has shortcomings as a validation criterion for the occurrence risk. We discussed this in 2.3.1, where we stated that we should control for the level of materiality and for the size of the audit sample. This, because the same error represents a greater risk with lower (stricter) materiality and with smaller sample size. So, when risk assessment is valid, for audit objects where the same level of materiality was chosen, the relation between risk assessment and error rate may be expected to be stronger than for the undivided organisations (see research question 5), because with constant materiality the size of the error rate has a direct bearing on the risk of a material error.
For the moment we disregard a similar effect of the sample size.

These considerations lead to the hypothesis implicit in research question 5: 'The correlation between error rate and risk assessment will be larger in groups of cases with the same materiality than in the set of all cases'. To test this hypothesis the level of materiality was categorized into 3 categories with boundaries 3% and 6%. For the cases in each of these categories the correlation was computed. Table 6.9 shows the results.

Table 6.9: P-correlation of OR x error rate for classes of materiality

Level of materiality	OR x error rate	size sub-sample
materiality<3%	.282	74
3%<=materiality<6%	.252	14
6%<=materiality	.516	29
total	.427	117

The table only shows an increase of the correlation at the highest levels of materiality; the increase has a p-value of .30, so is far from significant. In the two other classes, for 88 of the 117 cases (2 less than 119 because of missing values), the correlation decreases. This indicates that risk assessment is done with disregard of the level of materiality.

The introduction to this subsection implies that we should do a similar analysis for levels of the sample size. We choose to do this by transforming our data into the sampling risk (in the next section), which takes the sample size into account.

Conclusion research question 5: controlling for materiality increases correlations?
Contrary to the expectation for valid assessment of OR, the correlation of error rate and OR does not increase when controlled for levels of materiality.

Conclusion research question 2: validity with respect to error rate
Risk assessment and error rate correlate to a satisfactory degree, but the lack of monotony in the relation may cause problems.

Conclusion research question 10: varying validity with respect to error rate

The strength of the correlation varies too much over organizations as to justify the assumption that irrespective of organisation the same score on risk assessment implies the same extent of substantive testing.

Discussion
Eighteen percent of the variance in the error rate is explained when assuming a linear relation with OR, moreover the correlation is significant at the .0005 level. A satisfactory result. But care has to be taken for the nonlinearity, even non-monotony, that appears to be in the relation. The variation over organizations of the degree to which error rate and OR correlate, should have consequences for the way risk assessment is used in establishing the reliability of the auditor's opinion.

6.4 Risk assessment and sampling risk

6.4.1 Definition of sampling risk

In 1.2.4 we defined the sampling risk as:

> "Sampling risk arises from the possibility that the auditor's conclusion, based on a sample, may be different from the conclusion reached if the entire population were subjected to the same audit procedure." (the *sampling risk, SR,* see ISA 530 par. 7)".

In the audit methodology especially the risk of overlooking a material error is considered. This risk not only consists of the sampling risk (often referred to as the 'beta risk'), but also of the risk of analytical review and other non-sampling risks, like measurement errors (failing to see an error in a sampled transaction). Our validation criterion is based on the error as it was found in the sample. In chapter 4 we formulated a model which gave the possibility to calculate this sampling risk by way of a beta distribution. In our study we implemented this calculation by calculating the variable ESTRISK1, which gave the sampling risk SR as the upper-tail probability of the materiality in the appropriate beta distribution. For this distribution the most likely error was used as the mode; together with the sample size of the corresponding case the a- and b-parameter were assessed (see chapter 4 for the exact calculations). For some of the cases the most likely error was based on a judgmental sample. In such cases the auditor cannot give a statistical estimate of the error, but still there will be a sampling error. We calculated the associated sampling risk in the same way as for the statistical samples, so as to include the uncertainty about the estimate, due to all the non-sampled transactions, even if they were deemed to contain no error (almost surely).

The variable ESTRISK1 was classified into a variable POSTRIS1. So POSTRIS1 is a classification of the sampling risk ESTRISK1. The categories are as shown in table 6.10.

Table 6.10: Categories of sampling risk

category for sampling risk	boundaries
1: very low	SR < .05
2: low	.05 <= SR < .15
3: medium	.15 <= SR <.45
4: high	SR >=.45

The boundaries were chosen on the basis of a "statistical intuition"; we could not find real criteria for what should be "low", "medium", etc. Only the choice of 5 percent for the category "very low" is justified by the widespread use of this value for the significance level (the standard of at most 5% for the audit risk derives from the same

convention). Besides, a convenient consequence of the choice of these boundaries is that all classes of POSTRIS1 contain a satisfactory number of cases.

For an impression of how this categorisation works, we give a listing of some cases in table 6.11 in which the most likely error, the sample size, the level of materiality, the estimated sampling risk (ESTRISK1), and the category of sampling risk (POSTRIS1) are shown.

Table 6.11: Some values related to (category of) SR

MLE	sample size	materiality	ESTRISK1 (sampling risk, SR)	POSTRIS1 (category of SR)
,049500	46	,037	,80	4
,000000	59	,10	,00	1
,000000	58	,123	,00	1
,002980	113	,0275	,08	2
,001667	80	,0083	,57	4
,001580	63	,01	,57	4
,000000	187	,008	,22	3
,000008	30	,006	,83	4

Table 6.11 illustrates that the sampling risk naturally depends not only on the MLE, but also on sample size and materiality. So it meets the considerations in 6.3.3. In chapter two we argued that conceptually it is the most appropriate validation criterion. But its inclusion of the sample size also causes problems: errors of size 0, or very small errors, still can be associated with large sampling risks when the corresponding sample size is small. We will circumvent these problems by also creating the variable ESTRISK2, in the same way as ESTRISK1, but now by taking a constant sample size of 100 in the calculations. We will call this the "standardised sampling risk".

In the remainder of this section we will give the distribution of ESTRISK1 for categories of the occurrence risk, in the form of a scatterplot and also the correlation of occurrence risk, ESTRISK1 and ESTRISK2. Next we will give a cross tabulation and a graph of the occurrence risk against POSTRIS1. We will do this for the pooled organisations and also give results for the distinct organisations.

6.4.2 Results for the pooled organisations.

We start our analysis by creating a scatterplot for the sampling risk (ESTRISK1) against the occurrence risk (figure 6.6). The scatterplot shows a distribution of the sampling risks for all levels of occurrence risk which is considerably more homogeneous than in the case of the distribution of the error rates. As a consequence correlation analysis has more validity. Moreover, the picture shows that the distributions of the sampling risk at the levels of occurrence risk are relatively similar, so we should not expect a strong correlation.

Figure 6.6: Sampling risk by occurrence risk

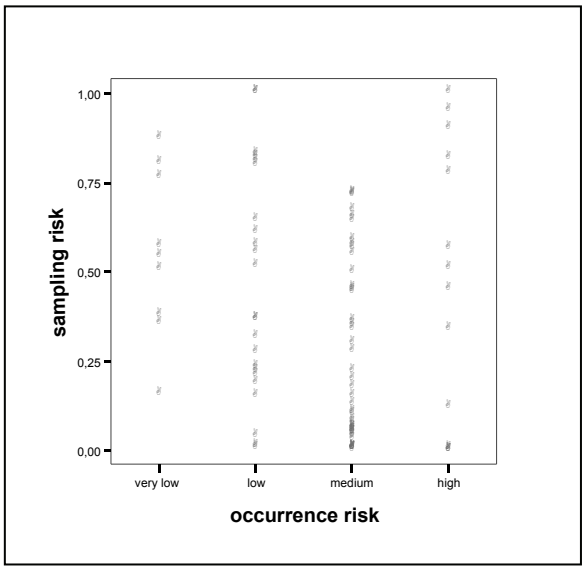


Table 6.12 gives these correlations. In this table, we also give the correlation of the standardised sampling risk (with sample size set at 100, ESTRISK2) with the occurrence risk. We give both the P- and the K-correlation

Table 6.12: Pearson and Kendall correlations of occurrence risk x sampling risk (ESTRISK1) and standardised sampling risk (ESTRISK2)

Organisation	P-correlation ESTRISK1	K-correlation ESTRISK1	n	P-correlation ESTRISK2	K-correlation ESTRISK2	n
all	-.179	-.179*	93	.05	0	117

*significant at the .05 level (2-tailed)

The varying number of cases (n) for which these correlations could be calculated are a consequence of missing values for the sample size. This problem was not met in the case of the standardised sampling risk because here we imputed a standard sample size of 100, also in the cases where the real sample size was unknown.

The table shows negative correlations, even to a significant extent in the case of the sampling risk itself and virtually 0 for the standardised sampling risk. For a good interpretation of these correlations we have to take the correlation between OR and sample size into account. This appears to be .101 (p-value .33). In 5.2.1 we argued that we might expect this correlation to be positive and that this positive correlation gives the correlation between sampling risk and OR a tendency to be negative. A correlation of .101 shows that in our study this tendency is only weak. This is confirmed by the correlations for ESTRISK2.

We also analyse the relation of a categorised sampling risk (cf. table 6.10) with the occurrence risk by way of a cross tabulation. This will give insight in lack of efficiency and lack of effectiveness, as can be caused by over- or under-assessment of the occurrence risk. Table 6.13 gives this tabulation.

Table 6.13: Cross tabulation of occurrence risk by categories of sampling risk

Sampling risk in categories→ Occurrence risk↓	Very low	Low	Medium	High	Row total
Very low			3 (33%)	6 (67%)	9 (100%)
Low	3 (13%)	1 (4%)	8 (35%)	11 (48%)	23 (100%)
Medium	14 (30%)	11 (23%)	10 (21%)	12 (26%)	47 (100%)
High	4 (29%)	1 (7%)	2 (14%)	7 (50%)	14 (100%)
Column total	21 (23%)	13 (14%)	23 (25%)	36 (39%)	93 (100%)

The cross tabulation indicates a negative relationship between occurrence risk and sampling risk. For a positive relationship the row percentages for an occurrence risk "very low" or "low" should show a decreasing trend, whilst the actual trend is increasing. For OR= "very low" the cells with values "very low" or "low" for the sampling risk in categories are even empty (while they should be the most filled). This is compensated more or less by some increasing trend in the row percentages for an occurrence risk "high". The fact that the Pearson Chi-square and the likelihood ratio test both are significant does not add to the relation looked for. The conclusion is that the negative correlations as given in table 6.12 are confirmed.

Ineffectiveness and inefficiency
A perfect relation between the sampling risk in categories (POSTRIS1) and the occurrence risk cannot be expected: even if it would exist it would be dependent on the choice of class boundaries for POSTRIS1. If it would exist, the table would only show zeros in the off-diagonal cells. So all positive counts in the off-diagonal cells indicate a less perfect relationship. More specifically: those in the above-diagonal triangle of the table correspond to cases where the actual sampling risk is higher than expected given the occurrence risk. Audits in these cases are designed based on an OR that should have been larger, which implies an audit of insufficient extent. So here a serious problem is detected: the audits have too large a probability to be ineffective. When we count all off-diagonal cases, there would be 40 cases with a threat of ineffectiveness. When, to account for the dependence on the classification of SR, we count cases at more distance from the diagonal there still remain (6+11+3=)20 cases, with a serious threat of ineffectiveness. A similar logic leads to the conclusion that 35 cases suffer from a threat of inefficiency, from which (4+14+1=)19 suffer a serious threat.

Remark: Ineffectiveness worse than inefficiency
From the viewpoint of validity ineffectiveness and inefficiency are of the same importance. But from the viewpoint of quality of the audit opinion, it must be stressed that ineffectiveness is a greater danger than inefficiency. Ineffectiveness may lead to an audit effort that is not sufficient and therefore to not detecting a material error, so to an audit opinion that lacks sufficient reliability. Inefficiency 'only' leads to an audit effort that is larger than necessary, so to an opinion that is safer than required.

The fact that no cases were found with the value '(very) low' for the sampling risk in categories when OR= 'very low', might be due to small sample sizes. This can be checked with the standardised sampling risk, classified in the same way, where a uniform sample size of 100 was used in the calculation of the sampling risk. Table 6.14 gives the results.

Table 6.14: Categories of standardised sampling risk by occurrence risk

Standardised sampling risk in categories→ Occurrence risk↓	Very low	Low	Medium	High	Row total
Very low	1	1	3	4	9
Low	7	3	19	6	35
Medium	15	6	20	16	57
High	5	0	2	9	16
Column total	28	10	44	35	117

The relation tends to be more in the direction as may be expected for a valid risk assessment. But at all levels of occurrence risk, deviations from this expectation occur. It can be seen, for instance, that only two cases appear for the two lowest classes of the sampling risk, and that the highest classes of the sampling risk still form a majority for OR="very low". In this way for every level of OR, there are deviations. So the correlation of virtually 0 as given in table 6.11 is confirmed in this analysis.

Ineffectiveness and inefficiency
The same logic as with the sampling risk (in table 6.13) gives us insight in ineffectiveness and inefficiency. We count 49 cases with a threat of ineffectiveness (in the above-diagonal triangle of table 6.14). When, to account for the dependence on the classification of SR, we only count cases at more distance from the diagonal there still remain (6+4+3=)13 cases, with a serious threat of ineffectiveness. A similar logic leads to the conclusion that 35 cases suffer from a threat of inefficiency, from which (5+15+0=)20 suffer a serious threat.

Conclusion research question 3: validity with respect to sampling risk
Risk assessment with respect to the sampling risk shows no validity for the pooled organisations; there is hardly any correlation between OR and SR.

Conclusion research question 6: OR strongest relation with SR as a criterion?
The answer to this question is evident: for the pooled organisations, the relation between error rate and OR was positive; between 'audit position' and OR the correlation was close to 0; a relation between SR and OR is negative or absent. The answer to this question is the opposite of what is expected for a valid risk assessment.

6.4.3 Results for the distinct organisations.

We did the same analyses for the distinct organisations, as we did for their combination. So firstly we calculated all P(earson)- and K(endall)-correlations for the six organisations that were our units of analysis. Table 6.15 shows the results. The differences in n for ESTRISK1 and ESTRISK2 are again caused by missing values for the size of the audit sample. So for instance in organisation 5 only in 4 cases the size of the audit sample was given, whereas in 21 cases an estimate of the error rate was available.

Table 6.15: Sampling risk (ESTRISK1) x OR and standardised sampling risk (ESTRISK2) x OR, by organization

Organisation	P-correlation ESTRISK1	K-correlation ESTRISK1	n	P-correlation ESTRISK2	K-correlation ESTRISK2	n
1	.104	-.09	12	.03	-.16	13
2	-.455*	-.342	21	-.05	-.102	22
4	-.165	-.114	15	-.251	-.162	20
5	.636	.548	4	.527*	.414*	21

Organisation	P-correlation ESTRISK1	K-correlation ESTRISK1	n	P-correlation ESTRISK2	K-correlation ESTRISK2	n
6	-.385	-.299	24	-.141	-.110	24
8	.756**	.485*	17	.772**	.507*	17
all	-.179	-.179*	93	.05	0	117

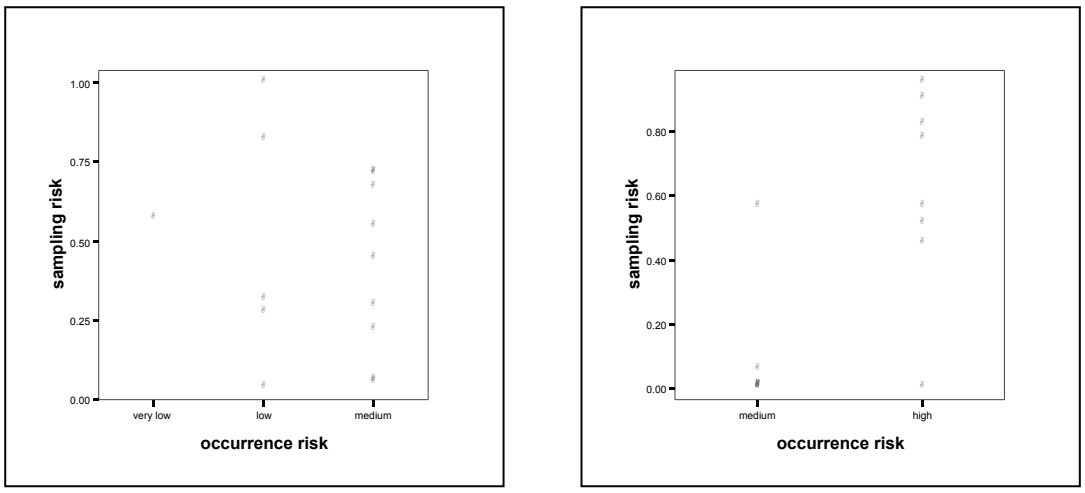
** significant at the .01 level (2-tailed) *significant at the .05 level (2-tailed)

Table 6.15 shows two organisations with a satisfactory correlation between OR and sampling risk. First of all organisation 8, where the correlations for both the sampling risk and the standardised sampling risk are significant; in the second place organisation 5, where the sampling risk shows high correlations, but which are not significant due to the small number (of 4) on which they are based. The standardised sampling risk is based on all 21 error rates that were available for organization 5. As this shows significant correlations with OR, organisation 5 can be deemed to have a satisfactory correlation. The change in magnitude of the correlations ESTRISK1, ESTRISK2, with OR, for organization 2 is remarkable: it changes from -.455 (and significant at the 5% level) to -.05. This is due to the sample sizes and materiality levels involved in this organization: there are relatively many cases with a relatively high level of materiality and correspondingly a relatively low sample size. With the constant sample size of 100 for ESTRISK2, the sampling risk for these cases decreases sharply. Obviously, this has a bearing on the size of the correlation coefficient.

We present scatterplots of the sampling risk by levels of the occurrence risk (ESTRISK1 by OR) for two organisations: again for organisation 4 and also for organisation 8. The latter to get some confirmation of the high correlations. Figure 6.7 gives the results.

The weak negative correlation for organisation 4 is confirmed: it does not depend on one outlier. The same applies to the scatterplot for organisation 8: this shows that the positive relation is not dependent on a single outlier. The quality of the correlation coefficients is also confirmed by the relatively small differences between the P- and the K- correlations.

Figure 6.7: Sampling risk by occurrence risk for organisation 4 and organization 8



The cross tabulation were also produced; table 6.16 gives the results for organisation 4

Table 6.16: Occurrence risk by categories of sampling risk for organisation 4

sampling risk in categories→ Occurrence risk↓	Very low	Low	Medium	High	Row total
Very low	0	0	0	1	1
Low	1	0	2	2	5
Medium	0	2	3	4	9
High	0	0	0	0	0
Column total	1	2	5	7	15

This analysis shows that for all levels of OR, where this can be concluded, the count of the classified sampling risk (POSTRIS1) increases for the higher scores of POSTRIS1. When the distribution of POSTRIS1 is approximately the same for the various levels of OR, the correlation cannot be far from 0. It is also remarkable that the only case for OR='very low' has the highest class of the sampling risk (with an audit sample size of 80); which certainly accounts for a part of the negative sign of the correlation. But figure 6.7 shows that the influence is only modest. The shaded areas in table 6.16 represent the (2+1=)3 cases with a serious threat of ineffectiveness and the 0 cases with the same for inefficiency.

Conclusion research question 3: validity with respect to sampling risk
Risk assessment with respect to the sampling risk SR shows validity for the distinct organisations for only two organisations (8 and 5). For the rest of the organisations validity is absent.

Conclusion research question 6: OR strongest relation with SR as a criterion?
For most organisations, a relation between SR and OR is either negative (with the unstandardised SR), or absent (with the standardised SR). So for most organisations the answer is negative. Only with organisations 5 and 8 the correlations of the sampling risk with OR are larger than that of the error rate with OR. But the difference is minimal.

Discussion on validity with respect to sampling risk.
We expected the validity with respect to the criterion variable 'sampling risk' to be stronger than with respect to the error rate, because the error rate has not the dimension of a risk, whereas the sampling risk does. But contrary to this expectation (for a valid risk assessment and in our study the correlation between sample size and OR being small), the validity with respect to the criterion sampling risk is virtually absent, with a majority of the correlations being negative where they should be positive. Only for two organizations (5 and 8), the correlations are satisfactory.

These findings suggest that an auditor's risk assessment applies more to the assessment of the error to be expected, than to the risk that the error might exceed the level of materiality. This conclusion is interesting to test in a next round in this research.

If this conclusion would hold in subsequent studies, it opens an interesting perspective on how to deal with risk assessment. Auditors could be asked to assess the most probable error in the audit population and also the less probable errors, up to the improbable and even impossible, ones. These assessments add up to a kind of distribution on the possible error rates and the probability in this distribution of error rates larger than materiality may be seen as the assessment of the occurrence risk, the risk of an unjustified unqualified opinion prior to the substantive part of the audit. We will recur to this discussion in chapter 10.

6.5 Risk assessment and conditional distribution of error rates

As stated in chapter 2, the distribution of the error rates, conditional on the assessed risk, can also be used as a validation criterion for risk assessment. In this section we show the results of this analysis only for the pooled organisations, because for the distinct organisations there are not enough data. We will give a scatterplot of the error rates for the four levels of OR and also the parameters of the beta distributions that can be fitted on the pooled data and on the data per level of assessed risk.

Figures 6.2 and 6.3 showed scatterplots of the error rates per level of the occurrence risk (OR). Because of the high density of datapoints near the OR-axis we needed a figure in which the error rate is treated as an ordinal variable (figure 6.3). But interpretation of the purely graphical representation is difficult. The only thing that seems to be clear is that the error rates tend to increase with the change to the level 'high' for OR: then error rates larger than 5% occur, where errors of this size are absent at the other levels of OR. The slight increase in error rates for the level 'medium' of OR is harder to take as an indication for an increasing tendency when changing from level 'low' to level 'medium'. Of course we are supported in this interpretation by the information of section 6.3, where we found that there is a positive correlation between the error rate and the occurrence risk. It also appeared that the means of the error rate per level of OR at first decrease ('very low'-> 'low') and next considerably increase.

To gain further insight than with these graphical representations, we will fit a beta distribution to the empirical distribution of the error rates. A beta distribution may be appropriate because it is defined for values between 0 and 1 (0, 1 included) and it is fit for modelling a large variety of distributions: from symmetric to highly asymmetric and from a very low to a very high dispersion (see for instance Novick & Jackson 1974). We will show that a symmetric distribution like the normal, does not fit at all. Once a (beta) distribution is fitted, differences in these distributions for the four levels of OR may become clearer, and as a consequence the degree to which these distributions indicate validity of the assessment of OR.

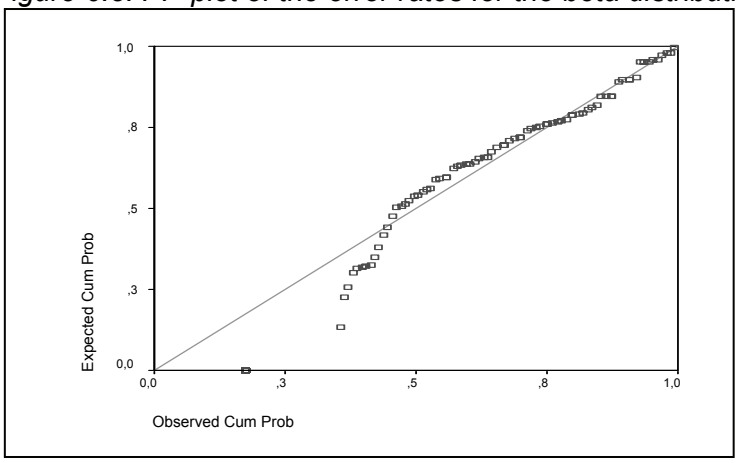
Before we try to fit a beta distribution on the data, it may be useful recall our discussion in 2.2.5 and 2.2.6, to note that the distribution of the 120 error rates, as they were found per case of an audit, is something different from the distribution of the possible error rates in a single audit case. The first, the distribution of errors per case, tells something about the probabilities that apply to a random selection from the set of the 120 audit cases in the research. It could also be used as a distribution for the probability of an error of a certain size, that applies for random selection from a set of annual accounts similar to the 120 audit cases. The second, the distribution of errors in a single audit case, applies to the distribution of the possible error rates in a single account under audit. In 2.2.7 we saw that this distribution can be seen as a property of the administrative processes, which directly relates to the 'real risk'. Prior to the substantive procedures, this risk is the occurrence risk. In 2.3.3 we saw that the sampling risk (also a single case outcome) can be seen as a direct indicator for this.

The fact that we model both distributions with the beta distribution does not take away this difference in meaning.

6.5.1 Results for the pooled organisations

We constructed a PP-plot[17] of the distribution of the error rates, for a beta distribution, and calculated the best fitting parameters, both for the total of error rates and for the error rates per level of the occurrence risk. We show the PP-plot for the pooled organizations and undivided for level of OR in figure 6.8.

Figure 6.8: PP-plot of the error rates for the beta distribution

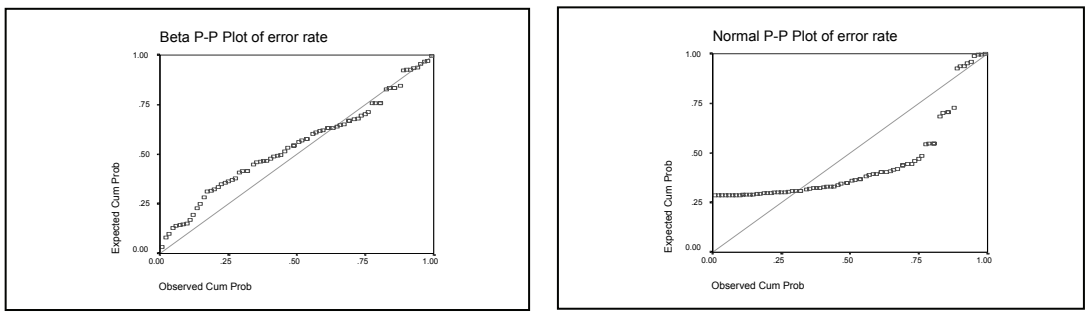


From figure 6.8 it can be seen that where some 35% of the observed error rates is smaller than the error rate corresponding to the first dot above the X-axis, in the fitted beta distribution values smaller than this observed value would have a probability of some 15%. So here the fitted distribution and the actual distribution do not fit. But the fit improves and is good for the values that are larger than the 45 percent smallest observed error rates

The poor fit for the lower cumulative probabilities is due to the 35% of the error rates equal to 0, where in a theoretical beta distribution none of any possible values has a probability larger than zero. The parameters of the fitted beta distribution turn out to be: a=.164 b=9.77.
This means that the mean of this distribution is (a/a+b=) .164/9.934=.0165. The difference with the actual mean of all error rates (.0167) is due to rounding precision: the fitted distribution is derived by the method of moments.

We further investigate the fit of the beta distribution by applying it to the non-zero errors and by also applying the normal distribution as a possible fitting distribution to the same non-zero errors. The results are shown in figure 6.9. The fit of the beta distribution has considerably improved. Now, for the lower cumulative frequencies, the expected frequencies are higher but the deviation is much less than for the error rates with the zeros included. The normal PP-plot clearly shows that the observed cumulative frequencies are far from symmetrical.

Figure 6.9: PP-plot non-zero errors for beta distribution and for normal distribution



[17] In a PP-plot the observed cumulative frequencies are plotted against the expected cumulative probabilities, as can be calculated from the model to which the data are fitted.

When we apply the same analysis to the error rate for the four levels of the occurrence risk (see also figure 6.2 for a scatterplot), the PP-plots roughly show the same fit of the beta distribution and the same evident lack of fit of the normal distribution. So again the beta distribution seems appropriate as a fitting distribution, but with some limitations for the lowest level of OR, as will be discussed next.

Table 6.17: Beta distributions for the errors by occurrence risk

Occurrence risk→ (Beta-) distribution↓	very low	low	Medium	high	all
N	9	35	58	17	120*
a-parameter	.186	.127	.388	.729	.164
b-parameter	25.34	38.71	38.18	9.40	9.77
mean	.00735	.00327	.0102	.0776	.0165

*in one case a score for OR was missing.

The means of the fitted distributions are equal to the means of the empirical distribution, as given in table 6.5, because of the fitting method. As shown in table 6.18, the relative rate of errors equal to 0 is high for the lower levels of OR. The rate clearly decreases with increasing OR. So one difference between the error distributions at the various levels of the occurrence risk is evident. But the fit of the beta distribution for the error rates larger than zero is influenced to a large extent by the zeros. Therefore we also fitted the distributions for the nonzero errors, for the levels ‘low’ and higher of the occurrence risk. For the level "very low" there are only three nonzero error rates, so fitting makes no sense. The results are shown in table 6.18.

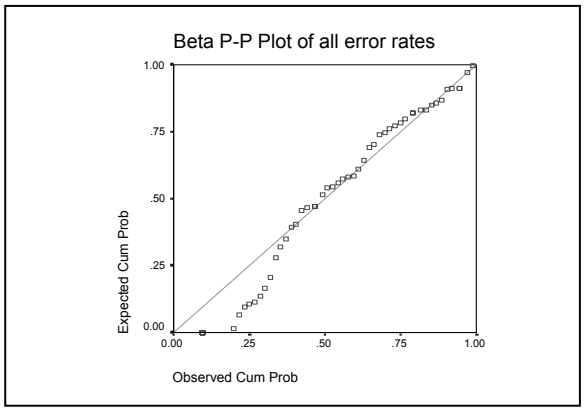
Table 6.18: Relative rates of 0 and beta distributions for the non-zero errors by occurrence risk

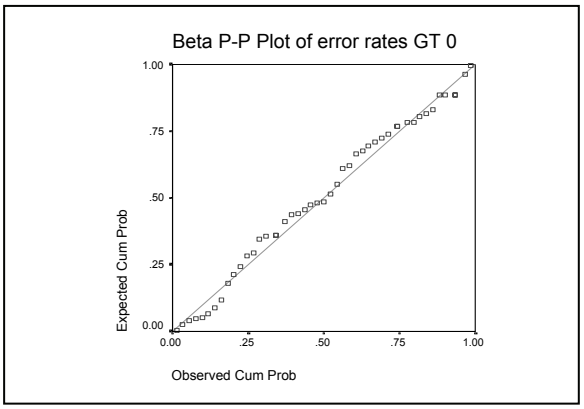
Occurrence risk→ (Beta-) distribution↓	very low	low	medium	high	all
#(greater than 0)	3	14	47	14	78
a-parameter		.378	.527	1.086	.280
b-parameter		46.057	41.944	11.35	10.699
mean non zeros	.0219	.00815	.0124	.0874	.0255
#(equal to 0)	6	21	11	3	42
rate of 0 (%)	.67	.60	.19	.18	.35
weighted mean	.0073	.00326	.0101	.0720	.0166

The monotonously decreasing relative rate of 0-errors, compared to the non-monotony of the relation of OR with the weighted means, may be seen as an indication that risk assessment has more predictive power for the existence of an error than for its size .

Just as with the fit of all cases, the fit of the beta distribution for the nonzero errors is much better than that with the zero-errors included. Figure 6.10 shows the fit by way of two PP-plots for the cases of OR=’medium’.

Figure 6.10: PP-plot for OR='medium' for all error rates and for the non-zero errors





It can be concluded that the conditional distributions shift in the expected direction for growing occurrence risk, taking into consideration both:

- the apparent predictive power for the existence of errors,
- the considerable difference in shape of the distributions, expressed by the decreasing b-parameter and the increasing means in table 6.18.

6.5.2 Results for the distinct organisations

There were not enough cases to perform this analysis on the level of the distinct organisations: it asks for an analysis at the 4 distinct levels of OR2, for 6 units of analysis. So on the average we would only have 120/24=5 cases to base the distributional analysis on. Therefore, we do not succeed in doing this analysis per organisation.

Conclusion research question 4: validity with respect to conditional distribution
Risk assessment with respect to the conditional distributions shows a satisfactory validity.

Discussion on validity with respect to conditional distribution.
In our discussion on research question 3 we speculated that the risk assessment of an auditor might sooner predict the size of the error to be expected in the audit object, than the actual risk of this error exceeding the level of materiality (the occurrence risk). The analysis with regard to the conditional distributions is consistent with this speculation: the analysis regards the distributions of the errors found, regardless the materiality that applied to the corresponding risk analysis. And still the conditional distributions show a satisfactory relation with the assessed risks. They extend our findings by showing that the predictive power of OR is best visible with the existence of an error, even stronger than with the size.

A promising perspective with regard to the conditional distributions is that they give the possibility to use purely statistical data (the error rates in a relevant set of financial accounts) in combination with risk assessment. We will go deeper into this possibility in the discussion at the end of this thesis in chapter 10

6.6 Moderator variables?

The strength of the relation between risk assessment and our validation criteria is found to be moderate, but with considerable variability over the criteria and over the

organisations. The strongest relations are found with the error rate and its empirical distribution as validation criterion, the weaker ones with the sampling risk. But also with the error rate as criterion, there is much variability: for some organisations the correlation is even virtually 0. It would be reassuring when the relevant relations would appear to be 'inflatable' by the use of convenient and interesting variables. So the question is: can we find moderator variables?

Potential moderator variables were already introduced in chapter 5, and formulated in research questions; we summarise:

- The complexity of the audit object, because with more complex objects, assessment is more complicated and therefore less valid risk assessment may be expected (see research question 7).
- The effort, put in the assessment, because with more effort, better assessment may be expected (see research question 8).
- The experience with the audited organisation: second or subsequent engagements might lead to more valid assessments than first engagements (see research question 9),
- The audit organisation itself, because the way in which risk assessment is dealt with will vary over organisations, which can lead to varying validity of risk assessment (see research question 10).

In section 6.3 we already found that indeed, seen from the level of 'all organisations', the organisation strongly acts as a moderator variable (see for instance table 6.3). In the next subsections we will investigate the other potential moderators mentioned above. For the sake of brevity, we will only do this with the error rate as the criterion variable.

6.6.1 The influence of complexity

We did not plan to measure complexity of the audit object. Therefore no direct measurement of this variable is included. But with one of the risk indicators, which will be discussed in the next chapter, the respondents were asked to give their opinion on the influence of complexity of the organisation on the occurrence risk. "Do you deem the complexity of the organisation to increase, decrease, or leave unaffected the risk of a material error?" was the question the respondent had to answer. This variable can be viewed as a proxy for complexity, because 'higher risk due to complexity' may be expected to coincide with 'more complexity' (in the opinion of the auditor). The fact that judgment on complexity is contaminated with judgment on the influence on risk, has to be kept in mind when interpreting the analysis with this proxy. Scores on this indicator could be given in three categories: risk decreasing (least complex), risk neutral or risk increasing (most complex).

Using this proxy as a classification variable, table 6.19 shows the results for the relevant correlations. The most complex organizations show the highest correlation for OR and the error rate, which is against the expectation underlying research question 7. Consistent with this expectation is the increase of the correlation with further decreasing complexity. When we take into consideration that there are only 14 cases in the class of most complexity, we could say that the expectation is at its best only weakly confirmed. But the differences found are far from significant: the largest, that between .546 and .363 has a p-value of .13 (computed by means of Fishers z-transformed P-correlations). So we must conclude that we find no difference.

Table 6.19: P-correlations over 3 classes of complexity

Complexity	OR x error rate	n	p-value difference me-le
most	. .616(*)	14	
medium	.363	65	.13
least	.546(**)	40	

** significant at the level of 0.01 * significant at the level of 0.05

Conclusion research question 7: is complexity a moderator?
The validity of risk assessment does not improve with decreasing complexity of the audit object. Research question 7 cannot be answered positively.

6.6.2 The influence of the effort

The effort was measured by way of a variable which stated the number of days spent in risk assessment. This variable was categorized in four approximately even numbered categories, as shown in table 6.20

Table 6.20: Categories of effort in assessment.

Category	Boundaries
1	effort at most 1.5 days
2	effort more than 1.5 days and at most 2.5 days
3	effort more than 2.5 days and at most 13 days
4	effort more than 13 days

With this categorization, the correlation between effort and the error rate behaved as shown in table 6.21. This table shows that the correlation is at its best when the effort category = 2 and almost equal for the effort category = 1 or 4. This pattern is inconsistent with the expectation underlying research question 8, which is that the correlation will increase with increasing effort. It is worthwhile to note that the differences for classes 1 and 2 (p-value 2%) and for classes 4 and 2 (p-value 1%) are significant at the 5 % level. But this does not confirm our expectation: more effort coincides with higher correlations.

Table 6.21: P-correlations over 4 classes of effort

Effort category	OR x error rate	n
Unknown	.333	8
1	.254	26
2	.703(**)	30
3	.479(*)	25
4	.214	30

** significant at the level of 0.01 * significant at the level of 0.05

Could there be an optimal effort?
Table 6.21 suggests that there might be an optimal effort. But, only when there is a convincing logic that implies the existence of such an optimum, it is justified to conclude accordingly. Looking for such a logic, we can imagine that, with increasing numbers of days spent, the load of information to be processed is growing at a rate as to cause errors in the information processing, causing a negative influence on the quality of the assessment that is larger than the positive effect of the extra information. But we think this 'logic' highly improbable, given the competence of auditors to deal with information streams of such sizes. We could not imagine other possible 'logics'.

Conclusion research question 8: is effort a moderator?
The validity of risk assessment does not improve with increasing effort spent in this activity. Research question 8 cannot be answered positively.

6.6.3 The influence of experience

As stated in the beginning of this section, we will investigate whether experience with the audit object positively influences the validity of risk assessment. In the available cases only 8 cases regard a first engagement. This makes it hard to make a proper comparison to the rest of the cases, because in so few a cases a relatively extreme outcome can easily happen. This should be kept in mind when seeing the results of table 6.23. But the fact that the 8 cases of a first engagement show a higher instead of a lower correlation between OR and error rate, surely is contradictory to the expectation that experience with the audit object improves risk assessment. (The difference of the correlations has a p-value of .10 (Fishers z-transformation)).

Table 6.23: P-correlations over 2 classes of experience

Experience	OR x error rate	n	p-value difference
first engagement	. .827(*)	8	.10
second and subsequent engagement	.530(**)	76	

** significant at the level of 0.01 * significant at the level of 0.05

Conclusion research question 9: is experience a moderator?
The validity of risk assessment does not improve with increasing experience with the audit object. Research question 9 cannot be answered positively.

Discussion on potential moderator variables.
None of the variables that we thought to be potential moderators turns out to be one, except 'organisation'. In the introduction of the relevant research questions (7 through 9) we already discussed reasons for a potential moderating effect. Are there more reasons for 'organisation' to be an exception, besides the reasons of (separate) cultures, mentioned in 5.2.3? We can speculate on possibilities like: differences in audit methodology, education, experience, audit objects. But our data do not provide for the possibility of testing these speculations. Research into organisation-related determinants of validity of risk assessment may nevertheless give relevant insight in possibilities to improve risk assessment.

6.7 Summary, Discussion and Conclusions

6.7.1 Summary

The outcomes of the analyses of this chapter are, to a certain extent, positive; namely:

(1) The correlations of the occurrence risk (OR) and the error rate on the average are positive and statistically significant.
(2) Also the distributions of the error rates for the various levels of OR shift into a direction that indicates valid risk assessment.

But there are also serious problems namely:
(1) We have used four variables as a criterion in this study and three of them show that validity suffers from problems. On one hand, the pooled data lead to the conclusion that risk assessment has explaining power for the criterion variables 'audit position' and 'error rate' (and its distribution). On the other hand, this relation is not stable: for some organisations it is strong and significant; but for others the relation is virtually absent. So we must conclude that there is too much variability over organisations. Without validating its own risk assessment, an organisation cannot be sure that this assessment is valid.

(2) Validity on the criterion 'sampling risk' is absent, also for its standardised form. Risk assessment shows no relation with, what can be seen as, the best (see 2.3.3) of the four validation criteria.
(3) Only the relation of the occurrence risk with the distributions of the error rate, conditional on the OR, shows no inconsistencies, but here we had insufficient data to investigate the relation at the level of the organisations.

6.7.2 Discussion

We discuss our findings in the form of some considerations with respect to the problems just stated.

The first consideration regards the variability in risk assessment between organisations; it was already discussed in 5.2.1. Here we must conclude that the evident variability between organisations causes the results of the analyses at the level of the pooled organisations not necessarily to hold for every distinct organisation. The outcome at this level could be interpreted as: "Auditors' assessments of the occurrence risk are valid with respect to the error rate", but not necessarily for every organisation.

A second consideration is that the anomalies might have to do with bias in the sample: only error prone cases have been collected. And maybe, in this type of audits, the assessment of risk is of lower quality than in 'error averse' cases. But, from the point of view of professionalism, you might expect better quality in error prone accounts. And even if bias in the sampling indeed prohibits full generalisation, the study still generalises to the implicit population of 'error prone' populations.

The limited possibility of generalisation was already discussed in the design chapter (chapter 5). We observed that the implicit population consists of error prone 'private accounts' and of a majority of the public accounts. This already means quite a population to be generalised to. Still a more general type of generalisation can be made: if the assessments of risks were valid, findings like in this study would (and should) not occur.

A third consideration is that the data could be of low quality. But all data have been checked and the nature of the data is quite simple; they could directly be got from the audit files, which makes erroneous answers quite unlikely. So the findings of this study are likely to be valid. With a view on the studies mentioned in 3.4.2.3, it still is not wise to take the conclusions simply for granted: it appears that outcomes of validation studies may vary, so before far-reaching conclusions are drawn from this study, it is better to replicate it in similar archival studies. In the mean time, where risk assessment can be invalid, it will be necessary not to rely on only this assessment, but to always do a significant piece of substantive work.

A fourth consideration regards the relatively loose way in which the risks are defined. That makes it hard to decide on how to validate risk assessment, or rather, unclear what has to be validated. Where validity is necessary, it is worthwhile to improve the possibilities of validation by defining more precisely the concept of occurrence risk. Now it specifies some risk of a material error, but that is all. It does not give any more information on what the auditor knows (or is able to assess) about the error rate in the audit population. Therefore neither of the four criteria used in this study is self-evident as validation criterion, although validity on these criteria still should be seen as a necessary condition for application of tables that transform risk assessment into audit effort.

A fifth consideration regards the finding that the relation between OR and error rate is relatively strong, both in the correlations and in the conditional distributions. It looks like risk assessment is best in trying to predict the size of the possible error. Even if this is something different from assessing a risk, it can be of use: assessing a 'most likely error', combined with assessments of less likely errors also can lead to risk assessment. We will discuss this to a fuller extent in the concluding chapter (10).

A sixth consideration regards an inconsistency between the practice of risk assessment and a basic assumption in the audit methodology. Audit methodology has an almost axiomatic assumption that the audit object will not contain a material error. And that therefore all the auditor has to do is to show that this is true. The only thing is that with a higher risk, assessed in risk assessment, he has to do more than at a lower level of risk. This in order to have sufficient assurance for his unqualified opinion. But at the same time risk assessment can very well produce a risk level of high. A naive person could think that this can not go together with the basic assumption of 'no material error' but it is audit practice that it does. It can only mean that risk levels of high still are maybe not larger than 10 or 20 percent. Speculating on their actual value does not make much sense; assessing this value by the analysis of the conditional distribution of the error rates, as indicated in this chapter, is much better. (Note that this still has to be completed by including the level of materiality).

6.7.3.Conclusion

These considerations lead to a (repeated) warning for the assessment practice: *Do not build on your assessments unless you know that they are valid*. In order to know this it will be necessary in many cases that validation be organised. In most audit practices this is not customary.

6.7.4 What next?

In chapter 5 we already stated that part of this study took a basic quality of risk assessment as a point of departure, which justified a plan to search for improvement of risk assessment. This basic quality seems to be present, be it with limitations, so a study in possibilities for improvement can take that basis. The improvement we try to establish will rest on the idea of decomposition of the assessment task. We will report findings on this idea in the next chapter.

It also seems appropriate to do something with the limitations. We should be alarmed by the unpredictability of the quality of risk assessment per organisation. We talk of "unpredictability", because when we would randomly draw one of the organisations that were in our study, we are not able to say beforehand if its risk assessment is valid or not, due to the considerable variability in the relation of risk assessment and the validation criteria. We may hope that decomposition of risk assessment will improve on this situation, but it seemed worthwhile to do a replication of the study of this chapter, in order to get a firmer basis for classical risk assessment. We did such a replication; in chapter 8 we will report our findings.

In many of the cases included in this part of our research, underpinning of the assessment of OR by system testing will have taken place. We could wonder why this did not lead to a more stable relation between our validation criteria and the occurrence risk. We did not collect separate data on system testing in these cases; therefore we cannot analyse its influence on the validity. So the question remains if consistent incorporation of system testing, as prescribed in many situations in risk analysis based

audits, improves the quality of the assessment of the occurrence risk. A study related to this question will be reported in chapter 9.

6.8 Summary of findings research questions 1 through 10

The next table gives an overview of the findings reported in this chapter.

Table 6.24: Overview of the findings in chapter 6

Research question	Result
1: To which degree will risk assessment (OR) correlate with the position of the error in a sample relative to the materiality?	Assessment of OR only shows insignificant correlations with the variable 'audit position'; it is not valid with respect to this variable (section 6.2).
2: To which degree will risk assessment correlate with the error rate in the audited account, for the pooled organisations?	Risk assessment and error rate correlate to a satisfactory degree, but the lack of monotony in the relation may cause problems (section 6.3).
3: To which degree will risk assessment (OR) correlate with the sampling risk (SR), for the pooled organisations?	There is hardly any correlation between OR and SR. (section 6.4)
4: To which degree will the distribution of the error rate vary with the level of assessed risk, for the pooled organisations?	The distribution of the error rate, conditional on the assessed level of risk, varies with risk assessment in the expected direction (6.5.1)
5: To which degree will the correlation between error rate and risk assessment increase when calculated for groups of accounts with the same level of materiality compared to the correlation for the whole group of accounts?	Contrary to the expectation, the correlation of error rate and OR does not increase when controlled for levels of materiality. (6.3.3)
6: To which degree will the occurrence risk (OR) show a relation with the sampling risk, which is stronger than that between OR and the 'audit position', or OR and the error rate?	For the pooled organisations: The answer to this question is negative or rather: the opposite of what is expected. For the distinct organisations: The answer to this question is negative or rather: in most cases the opposite of what is expected (and implied in the question). Only in two cases (with two organisations) the relation is as expected in a valid assessment. (section 6.4)
7: To which degree will the validity of risk assessment decrease with increasing complexity of the audit object?	The validity of risk assessment hardly improves with decreasing complexity of the audit object.(6.6.1)
8: To which degree will the level of validation increase with the effort put in the assessment of OR?	The validity of risk assessment does not improve with increasing effort spent in this activity. (6.6.2)
9: To which degree will the level of validation increase with the experience of the auditor with the business ?	The validity of risk assessment does not improve with increasing experience with the audit object. (6.6.3)
10: To which degree will the level of validation vary over organizations?	Validity with respect to 'audit position' varies considerably over organizations; it is significant for one organisation (section 6.2) The correlation between error rate and occurrence risk varies considerably over organisations (section 6.3). The correlation of SR and OR varies even stronger over the organisations than that of OR and error rate. (section 6.4). The distributional analysis was not done for distinct organisations, for shortage of cases.

Chapter 7: Risk Assessment by way of Risk Indicators.

Four relatively complete actions in three studies were taken in our research into the 'real risk': (1) validation of the classical assessment of OR, (2) an attempt to improve classical assessment of OR by decomposition of the assessment over risk indicators (1st and 2nd action formed the first study), (3) a replication of the validation of classical assessment of OR in the second study and (4) investigation of the predictive power of system tests for the error rate in the third study. Varying validity of classical risk assessment per organisation (1st action) and problematic results with risk indicators (2nd action) led to the choice for the replication (3rd action).
In this chapter we report the results of the 2nd action;.

7.1 Introduction

In chapter 6 we investigated the quality of classical risk assessment, in particular the assessment of occurrence risk. We found that the assessment of occurrence risk has positive qualities, but also some tricky anomalies. We also concluded that the plan of chapter 5 to look for improvement of risk assessment by way of decomposition seems useful, given the problems we met concerning risk assessment. We recall our conclusions of chapter 3 that decomposition of risk assessment may be expected to lead to improvement of the assessment and our introduction in chapter 5 of a decomposition of risk assessment by way of risk indicators.

In this chapter we will deal with the quality of risk assessment based on these indicators and consequently the possibilities they give for improvement, for the same organisations as in chapter 6. We will do this by answering research questions 11 through 16. Section 7.2 deals with research questions 11, 12 and 13, concerning the relation between risk indicators and classical risk assessment and other aspects of their consistency. Section 7.3 deals with research questions 14, 15 and 16, concerning the predictability of the error rate (7.3.1, 7.3.2). We also investigate the relations of transformed indicators with the error rate (7.3.3 and 7.3.4) and of the original indicators with a transformed error rate (7.3.5). In section 7.4 we discuss our findings and come to conclusions. In section 7.5 a table gives a summary of our results.

The data for this chapter were collected with the same questionnaire (see appendix 1), for the same organisations as in chapter 6. There was some variation over organisations in the indicators used; in this chapter we use the 19 indicators present in all questionnaires.

7.2 Risk indicators and the audit risk model

One of the desirable aspects of the risk indicators, is their consistency with the audit risk model (ARM). Risk indicators are hoped to improve risk assessment because the task to assess the occurrence risk (OR) is decomposed but, when the aspects represented by the risk indicators lead to assessments that are not related to the assessment of OR, acceptance of the indicators is only possible at the cost of rejecting the ARM. But the validity of this model has not been disproved so far. On the contrary: the findings of the previous chapter partially support its validity and, moreover, our

study of literature in chapter 3, shows that often risk assessment according to the ARM has desirable qualities such as sensitivity for experience and expertise and a relation with error rate (occurrence, size , or both). These observations are good reasons to aim at consistency with the ARM. (But still, a more consistent and stronger relation between risk assessment and error rate may be hoped to be achieved with the risk indicators, as a solution of the problems we also mentioned in chapters 3 and 6.)

We will investigate the consistency of the risk indicators and the OR by analysing their bivariate relations (7.2.1) and regressing OR on the risk indicators (7.2.2). In 7.2.3. we report outcomes of other checks on the consistency.

7.2.1 Bivariate relations of risk indicators and occurrence risk

The first thing to be expected in consistent relations are the bivariate correlations to be in the expected direction and of sufficient magnitude. We investigated these bivariate relations producing Kendall rank-correlations (K-correlations) between the occurrence risk and risk indicators and by producing cross tabulations.

Table 7.1 shows the K-correlations. The expected sign of the correlation is negative, because with OR a higher risk is represented by a higher score, whereas with the risk indicators a higher quality of the AO/IC is represented by a higher score, which in turn represents a lower risk.

Table 7.1: K-correlations risk indicator x occurrence risk for the pooled organisations (n=119)

risk indicator	K-correlation Risk indicator x OR	p-value
errors previous audits	-.307**	.0
changes controls since last audit	.118	.152
quality documentation admin procedures	.043	.6
separation duties EDP	-.286**	.001
segregation of duties other	-.048	.561
strength system of controls	-.334**	.0
access to EDP systems	-.234**	.005
access other systems and assets	-.160	.06
operation other segregations of duties	-.089	.289
are processes routine?	-.196*	.017
pressure for high-performance?	-.171*	.04
structural changes in organisation?	-.206*	.012
auditability of applicable regulations	-.010	.907
existence of audit trail	-.032	.698
nature of organisation	-.216**	.009
attitude of management	-.171*	.037
expertise of personnel	-.220**	.008
attitude of personnel	-.081	.323
complexity of organisation	-.016	.848

* significant at the .05 level (2-tailed) ** significant at the .01 level (2-tailed)

On the average the correlations between OR and the risk indicators are not high. But with consistency between OR and risk indicators, the bivariate relations cannot be

expected to be too strong, because variation on a risk indicator will only partially influence the occurrence risk. This because other aspects may be stable or even vary in a direction as to stabilise the occurrence risk or even let it vary in another direction. Still, 10 are significant (p-value <.05) or highly significant (p-value <.01) and only 2 out of 19 have the "wrong" sign. We applied the sign-test: it gives a lower-tail probability for the event (2 '–'signs in 19) of only .04%. We conclude a clear tendency for consistency of the risk indicators with the assessment of OR.

In order to check the nature of the relation for a relatively high and a relatively low K-correlation, the corresponding cross tabulations will be given; firstly for the strongest relation.

Table 7.2: Occurrence risk by strength of system of controls

controls→ occurrence risk↓	-1	0	1	row total
very low	2	2	5	9
low	1	13	21	35
medium	7	38	13	58
high	4	11	2	17
column total	14	64	41	119

Table 7.2 shows the expected relation: in the two lowest classes of OR the high 'strength of controls' are in a majority, and they are in a minority in the 'high' class of OR, where in OR 'medium' the risk neutral' class of 'strength of controls' forms a big majority. (Note that every cell in the cross tabulation is occupied even if this tabulation regards the strongest relation; this causes a scatterplot to give no insight in the relation.)

The next table represents one of the weakest relations, that of 'quality of documentation of administrative procedures'.

Table 7.3: Occurrence risk by quality documentation admin procedures

documentation → occurrence risk ↓	-1	0	1	row total
very low	2	3	4	9
low	7	16	12	35
medium	11	24	23	58
high	2	8	7	17
column total	22	51	46	119

Now for every class of the occurrence risk the distribution of the risk indicator is similar, which is fully consistent with a K-correlation of virtually zero.

We also show the bivariate relations for two distinct organisations: organisations 8 and 4. We chose these organisations because in the next subsection they will show to have the largest (8) respectively the smallest (4) explained variance in the occurrence risk, when we try to explain this by multiple risk indicators. Table 7.4 shows the results. For the sake of comparison we add the correlations for all organisations.

Table 7.4: K-correlations: risk indicator x occurrence risk for organisations 8 (n=17) and 4 (n=21)

risk indicator	K-correlation org. 8	K-correlation pooled organisations	K-correlation org. 4
errors previous audits	-.339	-.307**	-.164
changes since last audit	.070	.118	.242
quality documentation	.479*	.043	.124
segregation of duties EDP*	-.630**	-.286**	-.434*
segregation of duties other	.039	-.048	-.267
strength system of controls	.387	-.334**	-.273
access to edp systems*	-.373	-.234**	-.436*
access other systems and assets	-.310	-.160	-.302
operation other segregations of duties	-.635**	-.089	-.166
routine processing?	-.032	-.196*	.044
pressure for high performance?	-.350	-.171*	.131
structural changes?	-.354	-.206*	.144
auditability of law, regulations	.160	-.01	.027
existence audittrail	.149	-.032	.187
nature of organisation	.145	-.216**	.302
attitude management	-.462	-.171*	.070
expertise personnel	-.580*	-.220**	.279
attitude personnel	-.074	-.081	.153
complexity of organisation	-.267	-.016	.259

* Correlation is significant at the .05 level (2-tailed) ** Correlation is significant at the .01 level (2-tailed).

As can be seen, two of the correlations for organisation 8 are significant at the 1% level, two are significant at the 5% level, five are smaller (in absolute value) than those of the pooled organisations, so larger (absolute) values than those for the pooled organisations are a rule. Next to that, 7 correlations appear to be positive, the unexpected sign, when the occurrence risk and the risk indicators are assumed to be consistent. Even if none of the positive correlations is significant, this considerably reduces an overall picture of consistency.

For organisation 4 it turns out that the two significant correlations, that for 'segregation of duties EDP', and that for 'access to EDP systems', are the same as two of the five highly significant correlations for the pooled organisations. This is an indication of consistency. This is balanced by 12 of the 19 (non-significant) correlations being positive, which is highly inconsistent with the picture for the pooled organisations (and with the expected direction). Evidence and counter evidence make the consistency inconclusive.

Furthermore only one significant correlation for organisation 8 is also significant for organisation 4 and 15 out of 19 correlations for organisation 4 are smaller than the corresponding ones for organisation 8, moreover, 8 of 19 change sign . So the correlation structure for these two organisations is dissimilar. Obviously dissimilar correlation structures on aggregation will not add up to high correlations.

Conclusion research question 11: bivariate relations of occurrence risk and risk indicators.
For the distinct organisations, the bivariate relations neither give strong support, nor give counter evidence to consistency of risk assessment by means of the risk indicators on the one hand with risk assessment according to the ARM on the other hand.

For the pooled organisations the consistency emerges to a larger extent: almost all 19 correlations have the expected sign; 10 of them are significant, their magnitude is modest.

7.2.2 Multivariate relations of risk indicators and occurrence risk

With the bivariate relations being rather inconclusive as to the consistency of risk indicators and ARM, we will have to look at the multivariate relation for a more definitive answer to that question: can the occurrence risk be predicted from the risk indicators, or possibly a subset of them? When the indicators account for all the aspects that play a role in risk assessment in the classical sense, the indicators as a total, or a subset from them, may have a strong relation with the occurrence risk.

We analysed this by way of a stepwise regression analysis (which we started with the 9 'normative indicators' entered). This analysis omits the indicator that explains the smallest part of the variation in OR and starts a new round of analysis, in which similar actions are taken. Thus in every step an indicator is omitted, the least explaining first. This omission continues as long as it does not reduce the explained variance (of the regression model as a whole) too much. Actually in all cases in our analyses the R^2 decreased, but the adjusted R^2 always increased with the omission of the first predictors. This effect is due to the fact that the adjusted R^2 increases with decreasing number of predictors; and that this increase is stronger than the decrease due to omission of the predictors in the first steps. This also, because omission of a predictor only takes away the individual explaining power that is in the predictor and not the explaining power that is caused where it covaries with the other predictors that are left in the model. Obviously there is a point where omission of an extra predictor causes a decrease of the adjusted R^2. This effect gave us the opportunity to choose the solution with the largest explained variance for OR in terms of the adjusted R^2.

Stevens(1996) suggests to keep the number of predictors at less than $1/7^{th}$ of the sample size. Therefore we started our stepwise analysis with the limited number of 9 risk indicators. For the pooled organisations this number could have been larger, according to the criterion of Stevens, but for the single organisations it is still too large. However, as the stepwise analysis results in models with less predictors, in the end the models come close to, or satisfy Stevens' criterion. When in some cases the stepwise procedure still includes too many predictors, according to Stevens' criterion, we will do a check for overfitting of the model.

In the stepwise analysis the following 9 indicators were used as the predictors:
complexity of organisation
attitude management
segregations maintained?
errors previous audits
access to edp systems
strength system of controls
access other systems and assets
segregation of duties EDP
segregation of duties other

These indicators were chosen because they are deemed to refer to the central aspects in the assessment of risk (see for instance Arens and Loebbecke, 1997). In this thesis they will be referred to as 'normative' indicators. We took this approach to be preferable over a selection of indicators done by choosing the ones that correlate most with the dependent variable. Because, in that procedure, a capitalisation on chance almost inevitably will occur and it will lead to varying choice of indicators over the

organisations. This is less preferable, because a comparison of organisations is most informative when it can be done based on the same variables.
But there are also limitations:

- the information that is contained in the other indicators is not used in the analysis.
- we might have chosen indicators that do not represent the most important aspects in risk analysis.

We took these limitations for granted, because the outcomes of the analysis were satisfactory: the aim of the analysis, to get information on the consistency of risk assessment in the classical sense and by way of the risk indicators, was achieved. The same limitations will apply to our analyses with the error rate as the dependent variable (in section 7.3). But here we will do extra analyses with more predictors and no extra explaining power will be found. So here, the limitations are very mild.

The analysis led to the results given in tables 7.5 and 7.6. Table 7.5 gives the explained variance and its p-value for the distinct organisations (without the triplet, for its number of cases (13) was too small). In the analysis of the pooled organisations, the triplet was included.

Table 7.5: Occurrence risk explained by risk indicators; the explained variance

Organisation	Adjusted R^2	n	p-value
2	60%	22	.001
4	34%	21	.033
5	64%	22	.000
6	52%	25	.000
8	65%	17	.004
All	24%	119	.000

For an adjusted R^2 of 52% or higher the conclusion of this analysis can be that in the organisation concerned risk indicators can very well predict the occurrence risk. This means that the customary approach of risk assessment and the approach with indicators are highly consistent for these organisations. Only with organisation 4 the consistency is less, albeit that also in this case the regression model is significant.

For the pooled organisations the consistency is weaker. It is true that the model is highly significant (p-value: 0), but the explained variance is smallest of all. It looks as if the consistency per organisation is no guarantee for consistency at the same level for the pooled organisations. This might be caused by differences per organisation in the indicators that are most important in risk assessment, possibly because of the differing cultures in risk assessment. In this perspective, still, an explained variance of 24% means that there is a reasonable level of consistency between the assessments on the risk indicators and the assessment of the occurrence risk. In table 7.6, we show which indicators were included in the models for the various organisations.

Table 7.6: Occurrence risk explained by risk indicators; the included indicators

Organisation	Explained variance	variables in the model (in order of associated T-value; '+'-sign when positive regression weight)	significant predictors (at 5%)
2 ***	60%	5, 7, 2, 1, 8(+)	first 2
4	34%	1(+), 4, 5, 7	first 2
5*****	64%	2, 7, 1, 4, 8(+)	first 3
6	51%	4, 5(+), 6, 8	first 2
8****	65%	8, 4, 1, 6(+), 2	first 2
All	24%	4, 6, 8, 9(+), 2, 1	first 4

1: complexity of organisation 2: attitude management 3: segregations maintained? 4: errors previous audits 5: access to edp systems 6: strength system of controls 7: access other systems and assets 8: segregation of duties EDP 9: segregation of duties other

*** We checked this model for overfitting; with 3 predictors the explained variance (adjusted R^2) was 50%, p-value .001 **** idem; 3 predictors gave adjusted R^2 63%, p-value .001 ***** idem; 3 predictors gave adjusted R^2 60%, p-value .009 ; so if there is any, the overfitting is very limited

The analysis also shows which of the indicators are important in the assessment of the occurrence risk (OR). Table 7.6 shows that there is much variability in the most important indicators of the organisations. None of the indicators appears twice or more as the most important indicator. When we count how many indicators belong to the first two important in the models, 'errors in previous audits' appears three times, 'access to EDP systems' and 'access to other systems and assets' appear twice. So the 'errors in previous periods', scores the best with three, but disappears from the model for organization 2. Still, over organizations, this is the most consistently scoring indicator, which corresponds with its Kendall correlation for the pooled organisations (see Table 7.1). It is plausible that the variability in "most important risk indicators" is caused by the variability in the way risks are assessed in the various organisations, so that variability does not necessarily mean inconsistency. Our guess is, that the low explained variance for all organisations should be attributed to these varying assessment cultures.

Just as the correlations with the bivariate relations, in valid risk assessment the regression weights should be negative; because a higher score on the indicator represents more quality and a higher score on the occurrence risk represents more risk. Contrary to the expectation implied, some of the predictors showed positive weights; in table 7.6 a '+' is added after these indicators.

The '+'-signs mostly correspond to positive bivariate correlations. But it can also happen that indicators with a negative correlation with OR show a positive weight in the regression equation (e.g. 'access to EDP systems'). This can occur in regression analysis and is known as suppression effect (see Cohen and Cohen, 1983). With a view on this, no predictor can be left out of the model, once the indicators would really be used as predictors. Suppression is a possible consequence of collinearity.

Definition of collinearity:
When substantial correlations exist between predictors, these are said to be collinear. (Cohen and Cohen (1983)

Collinearity can cause difficulties with the computations of the regression equations and difficulties in the interpretation of a regression model. We will discuss this in more detail after table 7.10.

Conclusion research question 12: the multivariate relations of occurrence risk and risk indicators.

The multivariate relations per organisation are relatively strong, except for one organisation. This means that assessment on the risk indicators and classical risk assessment are consistent per organisation.
It appears that the regression models, that explain the assessment of OR from the nine 'normative' risk indicators vary strongly over organisations with respect to the risk indicators included in the model. This will cause the relatively low explained variance in the similar regression model for the pooled organisations. The findings can be interpreted as an indication of varying assessment cultures over organisations, given that the consistency per organisation is satisfactory.

7.2.3 More on the consistency of the risk indicators

In 5.2.3 we discussed desirable qualities of risk indicators. We reported what we did in order to construct the set of 23 possible risk indicators. These actions all were based on available files of completed audits and on the study of relevant literature. But an important source of evidence with respect to the consistency of the risk indicators can be added: the auditors of the organisations in which our research took place. Their opinion on the risk indicators, the extent to which they recognised their way of risk assessment in the risk indicators, can be seen as valuable information on the consistency of the risk indicators with audit practice. Making use of these sources of information, together with the previous analyses, implies a kind of triangulation (see Babbie, 1995) in our investigation of the consistency of the risk indicators. It results in an answer to research question 13 (in 5.2.4), as we now explain.

As we already indicated in 5.2.4, we took the opportunity to involve the auditors of the organisations with the introduction of our research in the organisation concerned. Here we discussed the quality of the set of risk indicators with our contact persons, together with some of the auditors that were to fill out the questionnaire. This discussion was on the qualities stated in section 5.2.3. In general our contact persons thought the set of indicators had the desired qualities. In some cases the discussion led to the addition of an extra indicator, which meant that the questionnaire for this organisation had a slightly different content compared to the others. The 19 indicators of this chapter form the cross-section for all organisations.

An additional check on the consistency of the indicators with the way auditors deal with risk analysis, was also introduced by inviting the respondents by way of an open question to mention factors that were of importance in their risk assessment. With some organisations almost all respondents took this opportunity. With others hardly any addition to the factors was proposed. In our analysis of the results of this invitation the question was answered if the factors mentioned should lead to the conclusion that a relevant risk dimension had been forgotten in the risk indicators. The answer to this question was negative. All factors could be interpreted as a specification or an alternative formulation of an indicator which was already in the questionnaire. Of course this is a subjective judgment, but within this almost inevitable constraint, the consistency of the indicators with risk assessment as viewed by the auditors themselves has been confirmed and the indicators can be seen as recognisable and exhaustive.

Conclusion research question 13: Can the risk indicators be seen as an appropriate representation of the view auditors have on audit risk assessment?
This consistency is confirmed by the judgment of the respondents that the indicators cover the way the respondent does risk assessment.

Our factor analysis in 7.3.3 will still give another approach to discus the quality of the risk indicators of this study. The discussion (in 7.3.3.2) will confirm the above conclusion.

7.3 Risk indicators and the error rate

With risk indicators that are consistent with the audit risk model, it makes sense to investigate the quality of the risk indicators as predictors for the error rate. It was the ultimate reason to introduce the risk indicators to improve on that prediction compared to the performance of the audit risk model, in particular its occurrence risk As argued in 5.2.3, this improvement can be measured by the extent to which prediction of the error rate is improved by this decomposition. In this section we will show the results of analysis on bivariate relations between risk indicators and the error rate (7.3.1) and of and a regression analysis of the error rate on the nine 'normative' risk indicators (7.3.2). In order to check if we omitted much information by focusing on the 'normative' risk indicators will also show the results of a regression analyses on factor scores (7.3.3) and of a regression analysis on scales constructed from the 17 risk indicators (7.3.4). Finally we regressed the error rate on a transformed error rate (7.3.5). We did the last three analyses only for the pooled organisations.

7.3.1 Bivariate relations of risk indicators and the error rate

To give insight into the relation between risk indicators and the error rate we will show scatterplots of the relation of a selection of risk indicators and the error rate; we will show the plots for the pooled organisations, for some organisations with high correlations and some with low correlations. As there is much more variation in the error rate than there is in the occurrence risk, we may expect these scatterplots to be informative. We will base our selection on the bivariate correlations, both the Pearson product moment- and the Kendall rank-correlation coefficient (P-correlation and K-correlation respectively).

Table 7.7 shows the bivariate correlations between the risk indicators and the error rate. It shows three series of correlations; the column with the P-correlation with one case omitted is discussed after the figures 7.1.

Table 7.7: Pearson- and Kendall-Correlations risk indicators with error rate (n = 119)

Risk indicators	P-correlation with error rate	P-correlation one case omitted***	K-correlation with error rate
errors previous audits	-.105	-.123	-.0137
changes since last audit	.125	.172	.023
quality documentation	.049	.085	.073
segregation of duties EDP	-.305**	-.275**	-.142
segregation of duties other	.091	.04	.042
strength system of controls	-.184*	-.093	-.184*
access to edp systems	-.313**	-.274**	-.113
access other systems and assets	-.012	.184*	.125
segregations maintained?	.033	-.033	.110
routine processing?	-.236**	-.189*	-.219**
pressure for high performance?	-.226*	-.208*	-.201**
structural changes?	-.086	-.243**	-.129
auditability of law, regulations	-.028	-.018	-.022

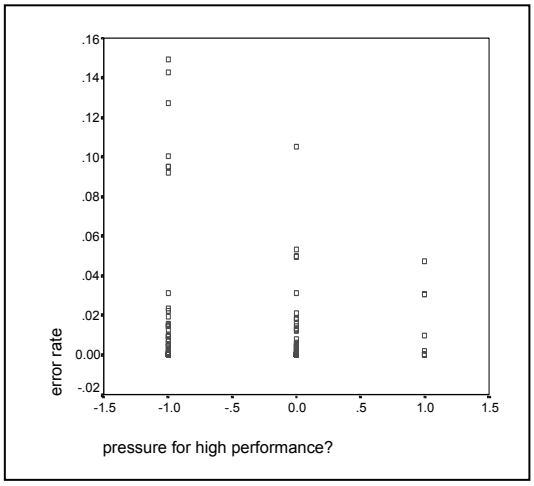
Risk indicators	P-correlation with error rate	P-correlation one case omitted***	K-correlation with error rate
existence audittrail	.058	.008	.007
nature of organisation	-.138	-.159	-.194**
attitude management	-.109	-.117	.031
expertise personnel	-.220*	-.254**	-.065
attitude personnel	-.053	-.042	.044
complexity of organisation	-.038	-.023	.016

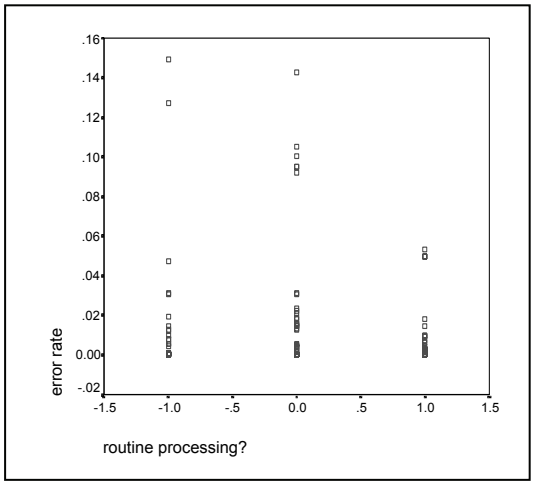
* Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).
*** The case with an extremely outlying error rate of .29 was omitted

The Pearson and the Kendall correlations are not inconsistent: most of them have the same sign and only the ones close to 0 differ in sign. But neither the picture given by the P-correlations nor the picture given by the K-correlations is satisfactory. As for the 19 P-correlations: six of them are significant and 14 of them, the significant ones included, have the expected minus sign. The sign test gives a p-value of .032 for at least 14 'minusses' out of 19 trials, so it can be concluded that the direction of the relation has a strong tendency in the expected direction. But the scatterplots in the figures 7.1 and 7.2 show that the P-correlations do depend on very irregular patterns, outliers included. So it is worthwhile to look at the K-correlations. According to these, the relation between risk indicators and the error rate is almost absent: only ten of the K-correlations have the expected minus sign; this has a p-value of exactly .50. This means that the only indications for a relation in the expected direction, are the 4 significant K-correlations having the expected minus sign.

Now in the figures 7.1 we show the scatterplots of 2 of the strongest correlations, with the P- and the K-correlation being consistent and in figures 7.2 of two of the almost zero-correlations. For the sake of the spread, in the figures 7.1 we left out the case with an error rate of 29%. This hardly biases the picture, because in both cases it coincides with a value of –1 for the indicator. So it deflates the correlation, but this still remains significant and negative (see table 7.7).

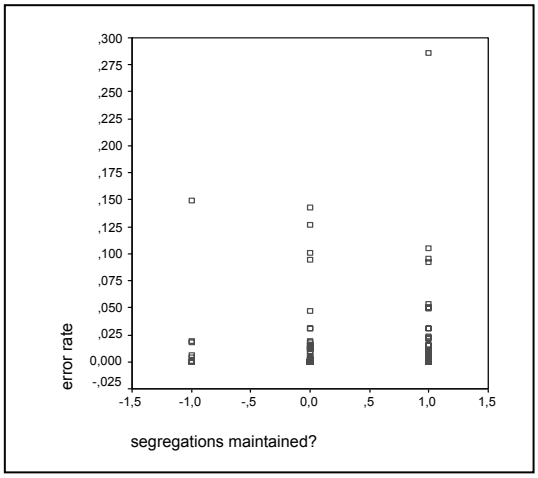
Figures 7.1: Scatterplots of two of the highest P-correlations:
error rate x pressure: - .226 *error rate x routine processing?:- .236*

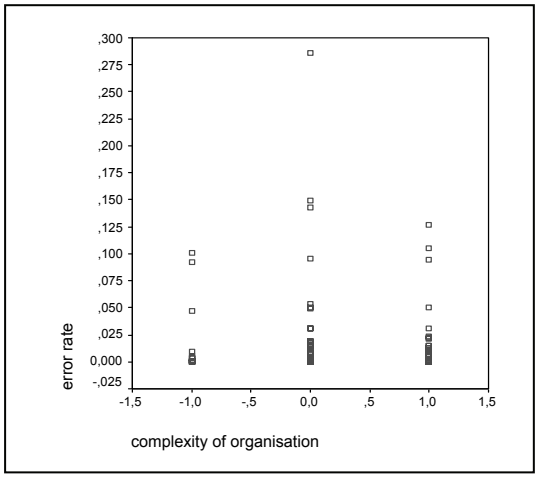




In the figures 7.2 we did not skip the outlier; the figures show how it affects the correlation for 'segregations maintained' (skipping changes the sign of the correlation, still both values are close to zero, see table 7.7) and to a lesser extent that for 'complexity organisation' (skipping only causes a small decrease in absolute value.)

Figures 7.2: Scatterplots of two almost zero P-correlations:
*error rate*segregations maintained:.033 error rate*complexity organisation: - .038*





Now with the distribution of the pairs (error rate, risk indicator) as in the scatterplots shown and with outliers of the size shown, correlation and regression analyses run a high risk of giving an unrealistic picture. Even rank-correlations run a risk of giving an unrealistic picture, because many of the values of the error rate are very close to each other. Given the variability of the error rate, this means that the rank order found in the error rates is an order which, to a large extent, is determined by chance.

Because it is a very obvious outlier, we calculated all correlations after skipping the case with an error rate of 29%. This affected all correlations, as is shown in table 7.7 in the middle column. But the number of negative P-correlations did not change.

Conclusion research question 14: the bivariate relations between risk indicators and error rate.
The bivariate relations are only very weak: there are some significant correlations, but the ratio of plus-signed correlations to minus-signed correlations is not convincing.

These considerations would lead to efforts to transform the indicators and the error rate in order to come to data that better meet the assumptions of statistical models. For the analysis of the multivariate relations we have chosen to do both: try and find the relations between the original indicators and the error rate (in 7.3.1 and 7.3.2) and relations between the transformed indicators and the error rate (in 7.3.3 and 7.3.4) and between the indicators and the transformed error rate (in 7.3.5). Because by doing both we can get an impression of relations that may exist between the raw data and we do an investigation into possibilities of finding better relations by transforming the data. We continue by investigating the multivariate relation between risk indicators and the error rate.

7.3.2 Regression of the error rate on risk indicators

In this section we will try to explain the error rate by way of linear regression analysis. Similar to the analyses of the risk indicators as predictors of the occurrence risk, we may expect the regression model to explain more variance in the error rate than is done by the distinct bivariate relations between risk indicators and the error rate.

We recall the nine ‘normative’ indicators we introduced in section 7.2.2.

1. complexity of organisation
2. attitude management
3. segregations maintained?

4. errors previous audits
5. access to edp systems
6. strength system of controls
7. access other systems and assets
8. segregation of duties EDP
9. segregation of duties other

Again we did a stepwise regression analysis for the pooled organisations and for the distinct organisations. The results are given in the tables 7.8 and 7.9

Table 7.8: Error rate explained by risk indicators; the explained variance

Organisation	Adjusted R^2	n	p-value
2	85%	22	0
4	45%	21	.011
5	23%	22	.084
6	10%	25	.119
8	32%	17	.094
All	22%	119	0

Two models for the distinct organisations have a significant p-value. The model for organisation 2 is even highly significant. For the remaining three organisations the models found are not significant. The explaining power of the models varies substantially over the organisations. Table 7.9 shows per organisation which indicators are included in the regression model.

Table 7.9: Error rate explained by risk indicators; the included indicators

Orga nisation	Explained variance	Indicators in the model (in order of associated T-value; '+'-sign when positive regression weight)	significant pre-dictors (at 5%)
2***	85%	9(+), 8, 7, 1, 4(+), 3(+), 6(+)	all
4	45%	2(+), 7(+), 6, 4	first 3
5	23%	2, 7, 1, 6	first 1
6	10%	6, 7(+)	none
8****	32%	5, 2, 7, 4, 1	first 1
All	22%	9(+), 8, 5, 6, 2, 7(+)	first 3

1: complexity of organisation 2: attitude management 3: segregations maintained? 4: errors previous audits 5: access to edp systems 6: strength system of controls 7: access other systems and assets 8: segregation of duties EDP 9: segregation of duties other

*** We checked this model for overfitting; with 3 predictors the explained variance (adjusted R^2) was 64%, p-value .000, **** idem: the explained variance (adjusted R^2) was 21%, p-value .113; so if there is any, the overfitting is very limited

When we look per organisation at the indicators in the two significant models the most striking observation is that various of the significant predictors have the unexpected ('+') sign. It appears that with 4 out of 6 indicators these unexpected signs also occur in the corresponding bivariate correlations, but with organisation 2 two of these corresponding correlations are negative, including the only significant one, as is shown in table 7.10.

Table 7.10: Error rate by risk indicators , P- correlations for the organisations 2 and 4

organisation 2 (n=22)			organisation 4 (n=21)		
indicator	P-correlation	p-value	indicator	P-correlation	p-value
6	-.424*	.05	2	.314	.17
9	.236	.29	7	.352	.12
3	.229	.31			
4	-.028	.9			

* Correlation is significant at the .05 level (2-tailed).

In our discussion of the results of table 7.6 we already introduced 'suppression effects' as a very likely cause for these illogical signs. The 'suppression' (see Cohen and Cohen (1983), Hair, 1998) comes from the introduction of an extra predictor variable (say pred 2), 'suppressing' variability in another predictor (say pred 1) that does not contribute to the variability of the dependent variable. This enlarges the explaining power of pred 1 and of the total regression model, but often leads to illogical signs for the regression weight of pred 2.

This means that the model relies on some interrelations between the predictors, that make it hard to interpret. This reliance is made visible when we remove one indicator from the model for organisation 2. When we remove 'segregation of duties other (9)', the most explaining, but with the 'wrong' sign, the explained variance dramatically drops from 85% to 38% (adjusted R^2); the associated p-value grows, but remains small: 3.3%. Now the most explaining indicator is 'segregations maintained?', but this predictor also has the 'wrong' '+'-sign. Apparently the model for organisation 2 is very sensitive to the changing of predictors and keeps relying on illogical signs for the regression coefficients. To a lesser extent this sensitivity also applies to the models for the other organisations, although they all show much less explained variance, which limits the effects of omitting predictors. As an example we modify the model for organisation 4, by omitting the indicator 'attitude management (2)'. Now almost all explained variance disappears: only 1.5% is left (adjusted R^2). At the same time the indicator "access other systems and assets (7)" keeps the 'wrong' sign.

The suppression effects obviously cause interpretation problems: can we say that an indicator explains much of the variability, when this explaining power is so dependent on other predictors. Next to these interpretation problems, collinearity may lead to unstable estimation. We will discuss this now.

Collinearity

In the discussion following table 7.6 of the illogical signs, we already noted that high collinearity also may influence the stability of a model in another way: when a regression model contains a predictor with a high multiple correlation with the other predictors in the model, the corresponding regression coefficient gets a high variance, making its estimate unstable. In order to look whether this collinearity occurs, we analysed all models for the error rate on collinearity, by means of the tolerance and the condition index. Tolerances larger than .10 and condition indices smaller than 15 are assumed to prevent this instability of the estimates (see Hair, 1998). Table 7.11 gives the (predictor with the) smallest tolerance and the largest condition index per regression model (as shown in table 7.9). The table shows that all diagnostic values are well within the range that should not cause (stability) problems.

Table 7.11 Collinearity diagnostics for regression models for the error rate**

organisation	adjusted R^2	#(predictors)	p-value (model)	Tolerance	condition ind.
2	85%	7	0	>.31	<5.3
2	64%	3	0	>.51	<3.4
4	45%	4	.011	>.55	<4.8
5	23%	4	.084	>.79	<3.9
8	32%	5	.094	>.71	<3.2
8	21%	3	.113	>.95	<1.5
all	22%	6	0	>.64	<2.8

**we omitted organisation 6 because the p-value of its model even exceeded 10%.

Conclusion research question 15: error rate predicted from the risk indicators?

Only for two organisations a model can be formulated that is significant at the 5% level and has more than 20% explaining power. But these two models are very sensitive to

the omission of one indicator, which makes them dependent on the interrelations of the indicators and not convincing for practical use.

Explaining power occurrence risk vs indicators
When we forget, for a moment, the interpretation problems with the models, we can look at the question whether the risk indicators give more explaining power to the assessments of an auditor than an assessment lead by the ARM. This regards research question 16.
Therefore we have to compare the variance explained by the bivariate relation to that by the multivariate relation. We do this in table 7.12 (ordered after adjusted R^2)

Table 7.12 Explained variance (error rate, indicators) x OR by organization

Organisation	P-correlation	R^2	adjusted R^2	n
6	-.009	0	10%	25
5	.518	27%	23%	21
8	.720	52%	32%	17
4	.121	1.4%	45%	21
2	.275	7.6%	85%	22
All	.427	18%	22%	119

For the explained variances R^2 (of error rate x OR) we refer to table 6.7; R^2 is the square of the P-correlation.
It can be seen that the explained variances per organisation increase in 3 cases and decrease in two cases. In two of the increase cases, the increase is dramatic, but unfortunately these increases are accounted for by models that we have shown to be hard to interpret. With 3 increases for the (adjusted) R^2, 2 decreases and an only very slight increase for the pooled organisations, we have to decide that it can not be concluded that the indicator approach leads to improved explaining power, compared to the classical ARM approach.

Conclusion research question 16: explaining power of risk indicators larger than that of OR?
No systematic increase of explaining power by the indicator approach could be shown, compared to the classical ARM approach.

7.3.3 Regression of the error rate on factorscores

As the suppression effects are caused by interrelations of the predictors, it is worthwhile to investigate the predictive power of the indicators, while eliminating interrelations. This can be done by applying a principal components analysis on the correlations of the indicators and regressing the error rate on the factorscores Because principal components are orthogonal, factorscores have zero correlation.

Another reason for using factorscores as predictors for the error rate is, that this circumvents the loss of information that may be caused by only using nine (the 'normative') indicators. A principal components analysis produces a number of principal components, each component being a linear combination of the scores on all indicators, so all indicators are included. Only a limited number of principal components is selected, such that the number of components is much smaller than the number of variables and such that the explained variance is satisfactory. The explained variance is the measure that accounts for the 'information content' of the principal components. So choice for a limited number of principal components also implies a loss of information, but this is not dependent on a choice of indicators.

The principal components can be used as new variables, also referred to as 'components'. Thus, as a result of the principal components analysis, every individual case has its scores on all indicators transformed into 'componentscores', more commonly referred to as 'factorscores'. These factorscores can be treated just like normal variables. We use this property by performing a regression analysis of the error rate on these factorscores, just as we did with the nine 'normative' indicators. It may be hoped that this analysis explains more variance in the error rate. This is not only because the factorscores contain information from all indicators, but also because the components solution maps the data on dimensions that are orthogonal, so any collinearity of the indicators that may affect the outcome of a regression analysis is excluded by the model itself. When these interrelations mask a relation between the error rate and the indicators, the relation with the factorscores may prove to be stronger.

But there may also be a decrease of explained variance, for several reasons.

1. The first may be the opposite effect of the argument just stated: a more sound data structure may also show that a relation between error rate and risk indicators is only weak.
2. The second reason stems from the fact that a principal components analysis is aimed at maximizing the explained variance in the indicators. This may result in a structure which not optimally predicts the error rate.
3. The third is that a components solution will always explain only a part of the total variance in the indicators.
4. This may be more so, because of a practical limitation: we can only analyse the data for the pooled organisations. None of the participating organisation has enough cases for a principal components analysis. An optimal result of the analysis asks for similar structures in the relations between the indicators over the organisations. When these structures are dissimilar, the factor solution will have limited explained variance and consequently have less predictive power.

With these considerations in mind we look at the principal components analysis.

7.3.3.1 Factorisation of risk indicators in seven components.

The analysis produced 7 principal components, which accounted for 67% of the variance. In this solution we included the components according to the common criterion: an eigenvalue larger than 1. Table 7.13 shows the outcome.

Table 7.13: Variance explained by principal components

Component	Initial Eigenvalues	% of Variance explained	Cumulative %
1	4,601	24,218	24,218
2	2,232	11,748	35,966
3	1,523	8,016	43,982
4	1,193	6,279	50,261
5	1,170	6,158	56,419
6	1,091	5,743	62,162
7	1,005	5,287	67,449

We rotated the principal components with a varimax rotation, which resulted in the matrix of factor loadings given in table 7.14.

Table 7.14: Rotated Component Matrix with factorloadings and interpretation*

risk indicator	Component						
	1	2	3	4	5	6	7
errors previous audits	,296	,424	,201	0	-,144	,327	0
changes since last audit	0	0	,195	0	-,795	0	,115
quality documentation	,415	,215	,299	0	0	-,593	0
segregation of duties EDP*	0	,131	,421	,402	,574	,200	,121
segregation of duties other	,460	-,159	,586	,306	,102	,101	0
strength system of controls	,182	,273	,568	,131	-,215	,155	0
access to edp systems*	,279	0	,291	-,253	,476	,479	0
access other systems and assets	,137	,156	,240	,248	0	,659	0
segregations maintained?	,202	0	,153	,833	,101	,109	0
routine processing?	,273	,197	,687	0	0	0	,263
pressure for high performance?	0	0	0	0	0	0	,894
structural changes?	-,106	,600	0	,268	0	0	,500
auditability of law, regulations	-,233	,596	,444	0	0	-,111	-,187
existence audittrail	,116	,718	0	0	0	0	0
nature of organisation	0	,789	,166	0	,253	0	,163
attitude management	,842	0	,206	,128	0	0	0
expertise personnel	,377	,463	0	,449	0	0	0
attitude personnel	,858	0	,233	,149	0	0	0
complexity of organisation	,515	,407	-,210	0	-,240	,321	0
Interpretation in short	attitude	nature organisation	nature job	seggregation	stability	access	pressure on employee

*All loadings smaller than .1 are rounded down to 0.

The distribution of the factor loadings over the indicators allows a relatively clear interpretation when a high threshold for the loadings is chosen. In table 7.14 we shaded the loadings larger than .59. These loadings determine our interpretation of the components, which is given in a keyword at the bottom of the column concerned. It must be noticed that with lower thresholds (for instance .40) for the loadings, the interpretability strongly reduces. This is no problem for the outcomes of the regression analyses: they are not affected, but it should be seen as a caveat when interpreting it. The interpretation of the components is given in table 7.15.

Table 7.15: Interpretation of components

Component	Interpretation
1 attitude factor	The indicators regarding the attitude with respect to internal control of the personnel and the management load highly on the first component; this is the 'attitude factor'
2 nature of organisation factor	Indicators regarding the 'state of the world' load highly on this component: the nature of the organisation, the auditability of laws and regulations, have there been changes. The existence of an audit trail not only refers to this 'state of the world' but also to the operation of controls; this is the 'nature of organisation factor' (with impact on auditability)
3 nature of job factor	The only indicator loading highly is that regarding the nature of the job, especially whether it has a routine character; interpretation is not selfevident, because indicators regarding controls and segregation of duties just miss the .59 criterion; still our interpretation is 'nature of job factor'

Component	Interpretation
4 segregation factor	The only indicator loading highly on this component is that regarding the operation of the segregation controls; the loadings of the two segregation indicators are relatively high, so support our interpretation: the 'segregation of duties factor'.
5 stability factor	Only one indicator loads highly on this component: that regarding possible changes in controls since the last audit; our interpretation: the 'stability factor'.
6 access factor	Two indicators load highly on this component: one regarding access control and one regarding the quality of the written documentation; our interpretation: the 'access control factor' supported by the relatively high loading of yet another access indicator.
7 pressure factor	Only one indicator loads highly on this component: that regarding pressure for high performance. So this obviously will be our interpretation of this factor

7.3.3.2 More on the quality of the risk indicators

We extend our discussion on the quality of the risk indicators (as announced in 7.2.3). The interpretation of the components shows that many of the relevant aspects of the object of risk analysis are covered. These relevant aspects can be divided into contextual aspects, the ones that cannot directly be influenced by organisational and control measures, and into organisational aspects: the things you can do to manage and keep control of the business- and administrative processes.

Contextual aspects are: (a) nature of business, (b) nature of organisation, (c) nature of job, (d) quality of personnel (expertise and attitude) and nature and stability of (e) environment and of the (f) organisation.

Organisational aspects are: (g) various types of segregation of duties, (h) various types of access control, (i) various controls to safeguard the quality of the processes.

The next table shows how they are related to the components.

Table 7.16: Coverage of aspects of risk assessment by the components

component→ aspect↓	Attitude (1)	Nature organisation(2)	Nature job (3)	Segregation (4)	Stability (5)	Access (6)	Pressure (7)
(a) nature of business,		x					
(b) nature of organisation		x					
(c) nature of job			x				x
(d) quality of personnel	x						
(e) stability of environment							
(f) stability of organisation					x		x
(g) segregation of duties				x			
(h) access control						x	
(i) various controls							

Only two of the aspects are not covered by the factor structure: the nature and stability of the environment and the various controls[18]. This can be seen as an indication that the structure and consequently the indicators cover the risk-assessment job of the auditor and have a similar structure. Moreover the indicators are exhaustive with regard to the relevant aspects of the assessment task, except for one contextual aspect. For only the stability of the environment is not included (it also is not present in the indicators); the aspect regarding the various controls has a fuzzy boundary with the other controls and moreover is present in an indicator. We conclude that the factor structure is very satisfactory: it allows a satisfactory interpretation and it covers the risk-assessment job to a large extent. Evidently this conclusion can be seen as a confirmation of the conclusion at the end of 7.2.3

7.3.3.3 The regression analysis

With this result of the principal components analysis, it makes sense to perform a regression analysis of the error rate on the factorscores. We did a simple analysis on the pooled organisations, in which we entered all seven factors. This led to a regression model with an adjusted R^2 of 11,6% and a p-value of .4%. This is only half of the explained variance found with the untransformed nine 'normative' indicators (22%; see table 7.8). One or more of the five possibilities mentioned at the start of this section will have caused this rather weak result. Table 7.17 accounts for one of these possible reasons: the models at the organisation level varied considerably. Moreover the plus-signs and minus-signs varied on the same factor for varying organizations, only with some exceptions: the factor "(segreg. of) duties" consistently had the '+'-sign, the factor "access" consistently had the '-'-sign. The other factors varied in sign. It should be noted that the expected sign again is the minus-sign. It also should be noted that the varying models over the organisations may account for the illogical signs.

Table 7.17: Error rate explained by factors, per organisation

Organisation	Adjusted R^2	components (sign of regression weight, if '+', in parentheses)	p-value	df
2	27%	Job nature, segreg. duties(+), stability, access, pressure	7%	21
4	43%	Attitude(+), organisation, stability(+), pressure(+)	1.4%	19
5	38%	Attitude, access	.5%	20
6	19%	Attitude, organisation(+), segreg. duties(+), access, pressure(+)	11%	24
8	63%	Attitude, organisation(+), stability(+), access, pressure	.5%	16
All	12%	all(1 factor '+'-sign, rest '-')	.4%	117

Table 7.18 gives the t-values of the coefficients of the factors in the regression model for the pooled organisations. Only two of the factors have a significant coefficient.

Table 7.18: Coefficients of the factors in the regression of the error rate

Component	t-value	p-value
attitude factor	-.743	.459
organisation factor	-.881	.380
job factor	-1.53	.129
segregation factor	.598	.551
stability factor	-2.552*	.012
access factor	-1.42	.158
pressure factor	-3.128**	.002

* significant at the 5% level (2-tailed) ** significant at the 1% level (2-tailed)

[18] When we observe that two segregation indicators and the 'strength of the system of controls' load highly on factor 3 we could also state that this lack of coverage is due to our strict cut off for the interpretation of the factors. We could also choose to lower the cut off value for the loadings (at the cost of loss of interpretability) and thus gain in coverage of the aspects.

Conclusion regression on factorscores
The explaining power of the factorscores is only half of that of the nine 'normative' indicators. The models for the distinct organisations strongly vary, indicating a lack of uniformity or consistency in risk assessment over organisations.

7.3.4 Regression of the error rate on risk scales

We constructed four scales, which we will call 'risk scales'. Each of the risk scales represented a major aspect of risk assessment: indicators that apply to properties

1. of the administrative system (SYSTEM, built with: segregation of duties edp, segregation of duties other, access other systems and assets, access to edp systems, segregations maintained?, strength system of controls;
2. related to the personnel (MANAPER built with: attitude management, attitude personnel, expertise personnel);
3. regarding the context of the audit object (POPULA built with: structural changes?, complexity of organisation, nature of organisation) and
4. regarding auditability of the object (AUDITA built with: auditability of law, regulations and existence audittrail).

The risk scales were formed on logical, substantive grounds: in principle the terms in a scale belong to the same kind of aspects: system indicators (segregation and access) combined with other system indicators, etc. Only indicators which caused the Cronbach's alpha of a scale to drop below .48 were skipped.

Constructing a scale out of various indicators has the effect of 'smoothing' the measurements. An extreme measurement on one indicator is likely to be compensated by a less extreme measurement on another in the same scale, unless the object itself is 'extreme' on that scale. This, as a consequence, also applies to the relations of the scales, especially with the error rate. A strong relation, caused by combinations of the predictors and the predictand that emerged by chance, is much less probable to occur, when the predictor is a scale. This may cause the relations found to change compared to that with the original indicators as predictors: the weak relations may grow (but not necessarily), the strong relations may shrimp (but not necessarily).

We performed a regression analysis of the error rate on these risk scales, in which analysis we also included the indicator for the expected influence of previous errors, because we expected it to have predictive value on its own that can not be missed. As in the previous analyses, we chose the model with the highest adjusted R^2. We show the results in tables 7.18 and 7.19.

Table 7.19: Error rate explained by four scales and one indicator; the explained variance

Organisation	Adjusted R^2 *	n	p-value
2	10%(85%)	22	.146
4	47%(45%)	21	.008
5	31%(24%)	22	.024
6	0 (10%)	25	
8	42%(32%)	17	.018
All	4%(22%)	118	.070

* variance explained by the nine 'normative' indicators in parentheses (see table 7.9)

Table 7.19 shows that for the organisations 4, 5 and 8 the explained variance increased. But for organisations 2, 6 and for the pooled organisations, the explained

variance (dramatically) dropped. Especially for the pooled organisations this is hard to explain, because the scales represent meaningful aspects of risk assessment and have good properties as a scale. For organisation 2 it may be related to the instability of the solution found, as discussed in 7.3.2 after table 7.8.

Table 7.20 shows which risk scales are included in the regression models

Table 7.20: Error rate explained by risk scales

Orga nisation	Explained variance	Indicators in the model (in order of associated T-value; '+'-sign when positive regression weight)	significant predictors (at 5%)
2	10%(85%)	system, popula (+)	none
4	47%(45%)	manaper(+), audita, erroprev, system	first 3
5	31%(24%)	manaper, popula, audita(+)	first 1
6	0 (10%)		none
8	42%(32%)	audita (+), system, erroprev	first 2
All	4%(22%)	system, popula, audita	first 1

Table 7.20 shows that even with scales significant predictors can have the unexpected '+'-sign in their weight. This strengthens the indication that some of the indicators (or a subset of them) are counter-intuitively related to the error rate. But this implies serious problems when the indicators in their present form would be used as predictors for the error rate. We will discuss these recurring illogical signs in greater detail in section 7.4.

7.3.5 Relation of the risk indicators with the transformed error rate

We concluded 7.3.1 with a discussion on necessity of transforming the error rate, because of the many outliers in the scatterplots shown in 7.3.1, combined with a great density of very small error rates and of zeros.
One of the ways to circumvent this problem is to categorise the values of the error rate. We did this by creating a variable "ERRCATEG" (category for error rate) as follows:

Table 7.21: Error rate categorised

value of error rate	Value of ERRCATEG
0	0
>0 and <.003	1
>=.003 and <.015	2
>=.15	3

With this transformed error we re-analysed three relations:
1. we recalculated the Pearson- and the Kendall-correlations,
2. we produced a cross tabulation of the indicator with highest correlation with the categorised error and one with a small correlation; these cross tabulations being an equivalent of a scatterplot
3. and we regressed the categorised error rate on the nine 'normative' indicators.

The first operation
The results of the first re-analysis are shown in table 7.22. For the sake of comparison we reproduce the P(earson)- and K(endall)-correlation with the untransformed error rate (in the 2nd and 4th column). We also shade the correlations of the nine 'normative' indicators with the (categories of the) error rate.

Table 7.22: P- and K-correlations risk indicators with categories error rate (n = 119)

P- and K- correlation with→ indicator↓	error rate (P-corr)	error category (P-corr)	error rate (K-corr)	error category (K-corr)
errors previous audits	-.105	-.152	-.0137	-.152
changes since last audit	.125	.074	.023	.048
quality documentation	.049	.031	.073	.044
segregation of duties EDP†‡	-.305**	-.195	-.142	-.162*
segregation of duties other	.091	.005	.042	.023
Strength system of controls	-.184*	-.222*	-.184*	-.208*
access to edp systems†	-.313**	-.151	-.113	-.112
Access other systems and assets	-.012	.128	.125	.114
segregations maintained?	.033	.091	.11	.109
routine processing?	-.236**	-.283**	-.219**	-.242**
pressure for high performance?	-.226*	-.216*	-.201**	-.207*
structural changes?	-.086	-.165	-.129	-.139
auditability of law, regulations	-.028	-.019	-.022	-.030
Existence audittrail	.058	-.043	.007	-.015
nature of organisation†	-.138	-.282*	-.194**	-.228**
attitude management	-.109	.015	.031	.021
expertise personnel†	-.220*	-.063	-.065	-.064
attitude personnel	-.053	.008	.044	.017
complexity of organisation	-.038	.005	.016	.011

*significant at the .05 level ** significant at the .01 level
† major change in P-correlation for error category ‡ major change in K-correlation for error category

We looked for 'major changes': changes from significant to insignificant or vice versa. As can be seen, according to this criterion the P-correlation shows a major change with four indicators (three as a decrease and one as an increase); only one of the K-correlations shows a major change (from insignificant into significant; the size of the change is not very dramatic). This result is not surprising. Some major changes may be expected in the P-correlation, as the categorisation of the error rate smoothes the outliers. So smaller P-correlations are to be expected. Only with exceptional covariation of error rate and an indicator, an increase of the P-correlation will occur. The changes in the K-correlation may be expected to be smaller, as the categorisation actually is a crude kind of ranking.

The 'normative' indicators covered only 3 of the 6 significant P-correlations with the error rate; they appear to cover only one of the four significant P-correlations with the categorised error rate. In bivariate relations for the pooled organisations, the 'normative' indicators appear to be not the best predictors.

The second operation
The second operation with the categorised error rate, showed some bivariate relation by a cross tabulation. We did this for the strongest and for one of the weakest P-correlations. Table 7.23 shows the result for the indicator concerning the complexity of the organisation, having a correlation of only 4% with the error rate.

Table 7.23 : Error category by complexity of organisation

error category→ complexity of organisation↓	0	1	2	3	row total
-1	2	6	3	3	14
	14%	43%	21%	21%	100%
0	28	10	10	17	65
	43%	15%	15%	26%	100%
1	11	10	11	8	40
	28%	25%	28%	20%	100%
column total	41	26	24	28	119
	35%	22%	20%	24%	100%

2-sided p-value of Pearson Chi-square: .120; 2-sided p-value of linear by linear association: .953

The tabulation fully confirms the absence of a relation: for all values of the indicator the row percentages do not show a systematic increase or decrease compared to the marginal percentages. The two-sided p-value of the Pearson Chi-square statistic is consistent with these observations: it is considerably more than .05.
Table 7.24 shows the result for the highest P-correlation.

Table 7.24: error category by 'routine processing?'

error category→ routine processing?↓	0	1	2	3	row total
-1	4	5	9	7	25
	16%	20%	36%	28%	100%
0	16	8	8	16	48
	33%	17%	17%	33%	100%
1	21	12	7	5	45
	47%	27%	16%	11%	100%
column total	41	25	24	28	118
	35%	21%	20%	24%	100%

2-sided p-value of Pearson Chi-square: .025; 2-sided p-value of linear by linear association: .002

Now for the low quality (risk high) score on the indicator, the row percentages show an increase for the higher error rates and for the high quality (low risk) score on the indicator a systematic decrease for the higher error rates, as may be expected. The p-value of the Chi-square statistic makes the observed relation significant at the 5% level. The linear by linear association is significant at the 1% level, consistent with the result shown in table 7.22.

The bivariate relations between indicators and categorised error rate are consistent with the corresponding bivariate relations for the untransformed error rate. This means that the picture of the bivariate relations with the untransformed error rate, should not be attributed to ill-behaving data, but should be seen as a fair representation of this (lack of) relation.

The third operation
The third operation consists of regressing the categorised error rate on the nine 'normative' indicators. The analysis produced a model with 5 predictors (indicators), having an explained variance (adjusted R^2) = 14,3%, with p-value = 0, n= 119. Table 7.25 shows the indicators that were included in the model, with the associated t-values.

Table 7.25: 'Normative' indicators predicting the categorised error rate

Indicator	t-value	p-value
segregation of duties EDP	-2.923	.004
access other systems and assets	2.795	.006
strength system of controls	-2.396	.018
segregations maintained?	1.999	.048
errors previous audits	-1.779	.078

Only the last predictor had a p-value larger than .05. Consequently four predictors have a significant regression weight. With two of them having the illogical plus-sign.

Regressing on the highest correlating indicators is a very inviting activity, because these indicators have both the logical (negative) correlation with the (categorised) error rate and also have much higher bivariate correlations. Moreover, a regression analysis on these highest correlating indicators might give an indication of possible improvement of the prediction of the error rate from better predictors. So we will make one exception on the policy chosen in which we only regressed on the 'normative' indicators, by doing one analysis with the indicators correlating highest on the categorised error rate. We chose the indicators with a correlation of at least .15 (in absolute value). From the models resulting from the stepwise analysis, we again chose the model with the highest explained variance (adjusted R^2). This gave a model with an explained variance (adjusted R^2) = 12.6%, with p-value = .001, n=119. Table 7.26 shows the indicators that were included in the model, with the associated t-values.

Table 7.26: Highest correlating indicators predicting the categorised error rate

Indicator	t-value	p-value
pressure for high performance?	-1.955	.053
nature of organisation	-1.886	.062
routine processing?	-1.473	.144
strength system of controls	-1.282	.202

None of these predictors was significant at the 5% level. All weights have the logical minus sign. This corresponds to the high negative correlations of the indicators with the categorised error rate. Except for the more logical model, with only negative weights for the predictors, the operation of taking the highest correlating indicators does not lead to more explained variance and none of the predictors is significant. So ruling out the illogical weights leads to insignificant predictors and less explained variance.

7.4 Summary, Discussion and Conclusions

7.4.1 Summary

The aim of this chapter is to investigate if risk assessment decomposed by way of risk indicators leads to an improvement of risk assessment. This investigation was done in three steps:

1. the consistency of risk assessment by way of risk indicators with risk assessment by way of the audit risk model was investigated;
2. the predictive power of risk assessment for the error rate, by way of nine 'normative' risk indicators, was investigated;
3. the predictive power for the error rate of linear combinations of these risk indicators was investigated.

The first step: that of consistency of the assessments by risk indicators and according to the audit risk model, led to satisfactory results. From this it can be concluded that the risk indicators cover the way auditors assess risks in their practice. This coverage was also found in interviews with practising auditors, in a pilot study analysing five audits (Broeze et al 1997), by comparison with the textbook of Arens and Loebbecke (1997, chapters 8 and 9), and in a logical analysis of the principal components. So we may conclude that the substance of the risk indicators was appropriate.

The second step: that of the predictive power of the "normative" indicators, led to varying results. The bivariate correlations between the risk indicators and the error rate in general are relatively small; regression analyses led to varying explained variances, which means that the predictive power is not stable over organisations. Moreover some regression models are very dependent on just one indicator. This, combined with many positive regression weights, where the logical sign is negative, makes the indicators as predictors of the error rate either not reliable or hard to interpret. Moreover the explained variance in the models with the risk indicators, compared with the variance explained by the occurrence risk, improved strongly in three cases but decreased in two cases. Although according to the first step, the substance of the indicators may be appropriate, the scoring was such that only three of the six highest correlating with the error rate and only one out of four of the highest correlating with the categorised error rate belonged to the nine 'normative' ones we selected. Our attempt to improve on the explained variance by regressing the categorised error rate on the eight highest correlating indicators led to a smaller explained variance

The third step: using linear combinations of risk indicators for the prediction of the error rate, neither improved their predictive power, nor eliminated the unexpected '+'-signs from the regression equations.

7.4.2 Discussion

It is hard to give clear causes for these disappointing results. The occurrence of the illogical positive regression weights especially is hard to understand, even if we take the phenomenon of 'suppression' (see 7.3.2, after table 7.10) into consideration. We discuss seven possible reasons for these results.

1. They may be due to the form of the risk indicators: we did not only ask for the quality of the aspects of the audit object that are relevant for the assessments task but, at the same time, we asked to assess the possible influence of these aspects on the risk of the occurrence of errors. Maybe that assessment interacted too much, and in a complicated way, with the assessments asked with the other risk indicators. Only decomposing into the quality of aspects like in Bell & Carcello (2000) might have been better (considering the positive result of Bell & Carcello).
2. They may be due to the indicators being formulated in a negative way: high-quality on the aspect represented by the indicator led to a high score on the indicator, but implied low risk, so a high score on a risk indicator would be expected to correspond to a low score on the error rate. Although this was explicitly explained in the questionnaire, confusion may have taken place when the respondents filled out the questionnaire: whether high-quality on the aspect covered by the indicator meant a high score or a low score. Although this is a possible cause, it would lead to systematically positive correlations. But in fact there are both (significant) positive and (significant) negative correlations.
3. They may be due to pure confusion, caused in some way or another. But then more or less random scores on all risk indicators should occur, which would

prevent significant results as we have found. Combined with a clear instruction in the questionnaire, this possibility is improbable.

4. They will be due to the phenomenon of 'suppression' as discussed in 7.3.2, after table 7.10. It will account for many of the illogical '+'-signs, thus offering a satisfactory explanation for these signs. But the difficulty in interpretation remains. Moreover, also bivariate correlations show the illogical '+'. So the suppression only partly gives a satisfactory explanation.
5. They may be due to the error rates being distributed in a very inhomogeneous way: many zeros, many small error rates a few larger error rates and some very large error rates. This explanation is partially satisfactory: we categorised the error rates into four size-classes. The regression analysis on this categorised error rate showed only 2/3rd of the explained variance of that with the untransformed error rate. This 2/3rd may be seen as the explanatory power that is robust against the ill-behaving distribution of the error rate. The 1/3rd that disappeared may be seen as distribution having its influence. Also in this analysis, the illogical regression weights do not disappear.
6. The anomalies appear to be organisation dependent as only some have the observed anomalies in the assessments of their auditors. Then this anomaly occurs for different indicators per organisation. Some (5 and 8) appear to have the logical relations in their assessments. To come to stronger conclusions, one could ask for larger samples per organisation, but the perspective is not very promising given all the illogical relations, per organisation, for the pooled organisations, for linear combinations of indicators and for the transformed error rate.
7. They may be due to susceptibility of auditors for illogical assessments. This possibility raises more problems than it solves, because if we would accept it, we would have to explain why this does not cause anomalies in a more consistent way. So we reject this possibility.

7.4.3 Conclusion

It must be concluded that the risk indicators, in the way we use them, do not systematically improve risk assessment compared to the assessment of the traditional occurrence risk.

7.4.4. What next

With so many problems and unexplainable anomalies, we decided not to continue the research on the risk indicators, but to continue the research by investigating the validity of the classical risk assessment of the occurrence risk, because the varying strength per organisation of the relation between error rate (and other criteria) and the occurrence risk raised questions concerning this validity.

We conclude with a summary of our findings, organised as answers to the research questions applying to the performance of the risk indicators.

7.5 Summary of findings research questions 11 through 16

The next table gives an overview of the findings reported in this chapter.

Table 7.27: Overview of findings chapter 7

Research question	Result
11: Are the bivariate relations between risk indicators and the occurrence risk in the expected direction and of sufficient strength?	For distinct organisations, the bivariate relations found neither form counter evidence, nor give strong support to the consistency of risk assessment by means of the risk indicators on the one hand with risk assessment according to the ARM on the other hand. For the pooled organisations, the consistency is more evident: almost all 19 correlations have the expected sign; their magnitude is modest.
12: Can the occurrence risk be predicted from the risk indicators by way of a regression model?	The multivariate relation for the pooled organisations is relatively weak, possibly due to dissimilar assessment structure over the organisations. The multivariate relations per organisation are relatively strong, except for one organisation. This means that assessment on the risk indicators and classical risk assessment are consistent per organisation.
13: Can the risk indicators be seen as an appropriate representation of the view auditors have on risk assessment?	Assessment of risk by way of the risk indicators and assessment of risk by way of the audit risk model are consistent. Moreover this consistency is confirmed by the judgment of the respondents that the indicators cover the way the respondent does risk assessment. Principal components analysis confirmed this consistency.
14: Are the bivariate relations between the risk indicators and the error rate in the expected direction and of sufficient strength?	The bivariate relations are only very weak: there are some significant correlations, but the ratio of (the expected) minus-signed correlations to (not-expected) plus-signed correlations is not convincing.
15: Can the error rate be predicted from the risk indicators by way of a regression model?	The indicators do not account for a convincing prediction of the error rate, neither with the original indicators as predictors, nor with predictors transformed into principal components or other scales, or with the error rate transformed into categories.
16: Is the explaining power for the error rate of the risk indicators larger than that of the classical risk assessment?	No increase of explaining power by the indicator approach could be shown, compared to the classical ARM approach.

Chapter 8: Validation of Risk Assessment Revisited

Four relatively complete actions in three studies were taken in our research into the 'real risk': (1) validation of the classical assessment of OR, (2) an attempt to improve classical assessment of OR by decomposition of the assessment over risk indicators (1st and 2nd action formed the first study), (3) a replication of the validation of classical assessment of OR in the second study and (4) investigation of the predictive power of system tests for the error rate in the third study. Varying validity of classical risk assessment per organisation (1st action) and problematic results with risk indicators (2nd action) led to the choice for the replication (3rd action).
In this chapter we report the results of the 3rd action.

8.1 Introduction

In this chapter we will perform a replication (as far as possible) of the research carried out in chapter 6. There we saw that there was much variability in the degree of validity between the organisations. In this introduction we will give an outline of research questions, design and other relevant aspects of this second study in our research. The validity of the assessed risk with respect to the 'audit position' (section 8.2), the error rate (section 8.3) the sampling risk (section 8.4) and the conditional distribution of the error rate (section 8.5) will be reported. We will do this both at the level of the pooled and the distinct organisations (where possible). In section 8.6 we will try and find some moderator variables and in section 8.7 we will look for the predictive qualities of the error 'of the previous year' for the error 'of this year'. In section 8.8 we will discuss our findings and come to conclusions. Section 8.9 gives a summary of the findings in the form of a table.

We concluded chapter 7 with the assessment that improvement of risk assessment by decomposition into the assessment on risk indicators did not work, at least not in the way we tried it, and that it would be wise to get a clear picture of the validity of classical risk assessment first. As a consequence of the conclusions of chapter 6 and chapter 7, we chose to continue this research in a second study, by doing another investigation into the validity of risk assessment in the classical way: the assessment of the occurrence risk in the audit risk approach. With this continuation, we hoped to find validity on the same criteria as used in chapter 6, but in a more stable way.

Participating organisations
In the second study, our data were collected from the audit files of the audit departments of the Dutch ministries, regarding the annual accounts of the year 2001. Files regarding the audit of these accounts are available to the Netherlands Court of Audit (NCA), as the natural consequence of its task to certify the annual accounts of all Ministries. For our research, we were granted access to these files, both by the NCA and the audit departments. For a start, all audit departments were a candidate for collecting the data from their files, but a kind of natural selection took place:

- firstly on the possibility of matching the assessed risks and the errors found without too much effort. The audit departments have there own logic in organising their audit files and as a consequence, it appeared that for many departments, the matching of risk assessment and related error would be too laborious;

- secondly on the necessity that only cases with "real risk assessments" were included. Some departments just set the OR at “high” in order to circumvent the necessity of testing the controls when assurance is derived from risk analysis. They deemed this to be more laborious than just doing substantive testing, or other substantive work. For our study these ‘assessments’ of OR are not useful, so these departments were not included in our research.

It appeared that at four departments (sometimes a part of) the necessary data, were available in the audit files: they regarded a complete audit risk approach, while error rate and assessed OR could be matched with reasonable effort. We will refer to them as “Organisation 1, 2, 3, 4”. None of these organisations also participated in the first study of this research. Also in this study, we agreed on the condition that the organisations would be kept anonymous.

8.1.1 The research questions

A selection of the research questions stated in 5.2.2 and 5.2.3 applies to this second study. We skipped research questions 7 and 9, because we did not have the necessary data on the complexity of the audit object and on first vs. repeated audits. Obviously, the research questions with respect to the risk indicators do not apply in this study. We summarise the research questions for this chapter.

Research question 1: To which degree will risk assessment correlate with the position of the error in a sample relative to the materiality?
Research question 2a: To which degree will risk assessment correlate with the error rate in the audited account?
Research question 2b: To which degree is this correlation stable over organisations?
Research question 3: To which degree will risk assessment correlate with the sampling risk (SR)?
Research question 4: To which degree will the distribution of the error rate vary with the level of assessed risk?
Research question 5: To which degree will the correlation between error rate and risk assessment increase when calculated for groups of accounts with the same level of materiality compared to the correlation for the whole group of accounts?
Research question 6: To which degree will the occurrence risk (OR) show a relation with the sampling risk, which is stronger than that between OR and the ‘audit position’, or OR and the error rate?
Research question 8: To which degree will the level of validation increase with the effort put in the assessment of OR?
Research question 10: To which degree will the level of validity vary over organisations?

We added two new research questions:

Research question 18: To which degree does the validity of risk assessment depend on the type of transactions involved?
Our data regarded two types of transactions: obligations and payments. These are closely related, but in practice their accounts are separately analysed as to the occurrence risk that applies to them. It is hard to tell whether the type will influence the validity of the assessment of the occurrence risk. Obligations are at the basis of every payment; especially legality is accounted for when the obligation is made. And in governmental transactions legality is a very important feature, so it may be expected that risk assessment with respect to obligations will be done very carefully. On the other hand, the real expenditure is done in the payment. Thus an error in the payment has direct financial consequences, so it may be expected that risk assessment with

respect to payments will also be done very carefully. It is hard to say which of the two reasons for 'extra care' will win, so we chose to formulate question 17 as an open question.

Research question 19: To which degree will the error of the previous year predict the error for this year?
By including this research question, we investigated the possibility that the error of the previous year may be a valuable extra in risk assessment, or even a replacement for risk assessment if it would turn out to be invalid. When the administrative processes are stable from last year into the current year, it hardly needs an explanation that the error of the previous year may be expected to have this predictive quality. Therefore in risk assessment the error of the previous year is always an important feature of the audit object. But we also have an opposite possible effect, discussed in 3.2.1 leading to expectation E, that the error of this year would regress negatively on the error of last year. So we can also say that by including this research question we investigated a part of the basis of risk assessment. It opens the possibility of a comparison of the error of previous year with risk assessment on their performance as a predictor of the error rate (of this year). Unfortunately we do not have data on the possible changes in the administrative processes, making it impossible to test for either the stability or the negative regression effect.

8.1.2 The data

The data were collected by means of a questionnaire that again was discussed with the audit organisation that was willing to provide for the data. The discussion was less complicated than that in the first study, because no risk indicators were involved. The 'questions' were only meant to guide the data collection from the audit files, which we ourselves did in this study. Two other employees of the NCA were also involved in the data collection and in the discussions on the analysis. The data regarded the annual accounts of 2001.

The number of audit cases per selected organisation varied considerably. With three organisations (2,3,4) we used all available audit cases, satisfying our information need; with one organisation (1) we used 37 of the available 70 cases. The degree of detail and the information available on a number of background variables also varied. Background variables were chosen, which could possibly serve as moderator variables. Per audit case we collected the following data (if available):

- the nature of the transaction (payment or obligation)
- is it a first engagement (yes or no)
- the number of days spent for risk analysis
- the nature of the transaction (routinely or not)
- the level of materiality
- the size of the account (in Dutch guilders)
- the way transactions were selected (by statistical sample, or other)
- the size of the sample
- the inherent risk
- the internal control risk (results of system testing included)
- the size of the error found (before corrections, if applicable)
- the audit opinion.

8.1.3 Data processing

In this second study of the research, we had the information on the inherent risk (IR) and the internal control risk (ICR). Our research questions were on the occurrence risk

(OR). So we had to transform IR and ICR into OR. We did that by making use of the table of HCDAD (1997) in which for every combination of IR and ICR a level of assurance is given that has to be generated by the substantive testing, or other substantive procedures (see section 1.1). The following table shows this transformation.

Table 8.1: Transformation of inherent risk and internal control risk into "reliability lack"

Inherent risk→ Internal control risk↓	low	medium	high
low	.67	.78	.83
medium	.83	.89	.92
high	.90	.93	.95

In this thesis we call the assurance to be generated "reliability lack". Obviously it represents the assessed occurrence risk. In some analyses we will categorise this 'reliability lack' in 5 categories, by means of a new variable OR2, created from inherent risk and internal control risk as shown in table 8.2.

Table 8.2: Transformation of inherent risk and internal control risk into "OR2"

Inherent risk→ Internal control risk↓	low	medium	high
low	3	4	5
medium	4	5	6
high	5	6	7

Because in the data OR2=4 (which corresponds with two possible values of 'reliability lack') does not occur, for every value of the "reliability lack" there is exactly one value of OR2 (but not v.v.). In this chapter we will use the names "reliability lack (occurrence risk)" for the transformation of table 8.1 and "occurrence risk" for the transformation of table 8.2.

Organisation 2 did not assess the inherent risk, but only the internal control risk. For the analysis of the pooled data, we set the inherent risk at "high", so that risk assessment for this organisation became compatible to a certain extent with that of the other organisations. But by imputing a standard value for the inherent risk, we influence the outcomes. Therefore we will never give a result for the pooled organisations, organisation 2 included, without also giving the result for the pooled organisations, organisation 2 excluded and in various analyses we only pool organisations 1,3 and 4.

8.1.4 Validity and generalisability

We got data from four audit departments of the Dutch government. The same considerations on generalisability as in the first study apply to this second study: the sample of organisations is not random, but a kind of "convenience sample"; the conclusions apply to the organisations in the sample, not to the individual auditors.

Also in this study, it is not clear to which extent the conclusions for the participating organisations apply to risk assessment in general. Actually, this also would be true if the sample of participating organisations would have been random from some clearly defined set of organisations. This, because a size of 4 hardly allows interesting and statistically sound conclusions to whichever larger set. So again, as in the first study, we show the existence of phenomena and can state that they apply to some wider set of organisations, but we cannot give clear boundaries for this set.

8.2 Risk assessment and 'audit position'

8.2.1 Definition of 'audit position'

For this analysis we again use the variable "audit position" as we defined it in 2.3.2. It is "OK" when the most likely error is smaller than materiality. It is "not OK" in the other cases. In case of a valid risk assessment, the "not OK"-positions may be expected to coincide with the higher "reliability lack" (occurrence risk), so when "OK" is coded with 1 and "not OK" with 0, a negative point-biserial correlation may be expected. Again we note that "OK" needs not be the same as 'acceptable' for in general, the auditor comes to his audit opinion on sharper criteria.

8.2.2 Results for the pooled organisations

Table 8.3 gives the results of a cross tabulation of "reliability lack" (occurrence risk) with "audit position". Unfortunately the analysis is based on only 2 "not OK"-positions. This makes the results very dependent on the position of just two cases, something to be kept in mind when interpreting the outcome. This shows only a weak a relation: the two "not OK"-positions are only just above the middle of the range of the assessed risk ('reliability lack') and not in the highest range, as may be expected in valid risk assessment.

We also calculated the point-biserial (PB) correlation; its value was -.075, with a p-value of .60. This confirms the interpretation of the cross tabulation.

Table 8.3: Reliability lack (occurrence risk) by 'audit position' (pooled organisations)

'audit position'→ Reliability lack↓	Not OK	OK	Row totals
.67	0	6	6
.83	0	4	4
.89	0	12	12
.90	1	2	3
.92	0	4	4
.93	1	14	15
.95	0	8	8
Column totals	2	50	52

(PB-correlation: -.075, p-value: .60)

In table 8.3 organisation 2 was not included. When we do this (with, as known, the inherent risk by default set at "high"), the analysis is based on 76 cases, the number of "not OK"-positions remains only 2, the point-biserial correlation becomes -.103 with a p-value of .38. For the pooled organisations, the conclusions can only be that risk assessment shows no validity with respect to "audit position".

8.2.3 Results for the distinct organisations

We applied the same analyses to the distinct organisations. The two "not OK"-positions occurred in organisation 1. This means that for the other organisations the "audit position" will be constant; when there is no variation, 'co-variation' is meaningless and so is correlation. So for the "audit position" as a criterion, there is only one organisation for which the analysis of validity makes sense. The following table (8.4) gives the result of the K- and the P-correlation.

Table 8.4: K- and PB-correlations 'audit position' x reliability lack (occurrence risk) for organisation 1

Organisation	K-correlation	p-value	PB-correlation	p-value	n
1	.033	.83	.052	.76	37

Conclusion research question 1: validity with respect to 'audit position':
Assessment of the occurrence risk only shows correlations close to 0 for the pooled and the distinct organisations. Validity cannot be concluded.

Obviously, research question 10, with respect to the variability of validity over organisations, cannot be answered.

Discussion
It should be kept in mind that the conclusion is based on only two cases with "not OK"-positions. So the empirical basis of the conclusion for this research question is not very strong. We could consider a change in the definition of 'audit position', by only calling it OK when the upper tail probability of materiality, given the error rate, is smaller than, say, 5%. But actually we will go in this direction when taking the sampling risk as a criterion, in section 8.4. So we leave it the way it is, noting that apparently the criterion 'audit position' is crude and therefore, with these data, not very informative on validity.

8.3 Risk assessment and error rate

8.3.1 The error rate

For this analysis we used the variable error rate, as we introduced it in 2.3.1. It was directly found from the audit files and represented as a percentage of the account-size.

We started section 6.3 by stating

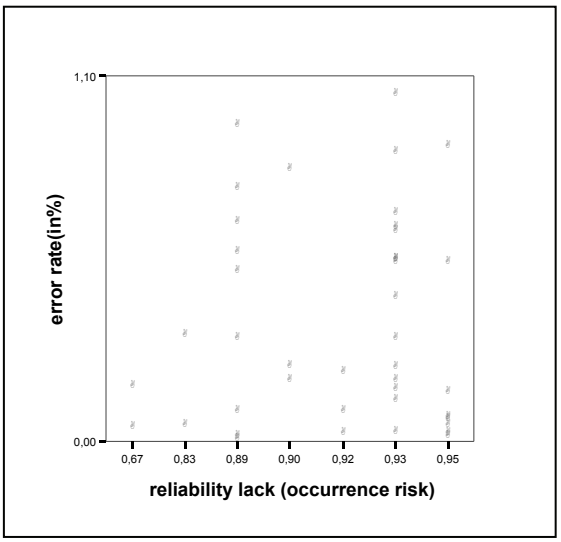
- that the error rate itself is an indicator for the level of risk associated with the administrative processes,
- that it should be kept in mind that the occurrence risk is ordinal by character, which makes the Pearson correlation less fit as an association measure,
- that therefore a rank correlation is more appropriate;
- and that we should be aware that the error rate itself may have a pathological distribution: very inhomogeneous, with extreme outliers.

These remarks also apply to this section, with "reliability lack" as a measure for the occurrence risk. Extreme outliers were actually found in chapter 6, which adds to the relevance of the scatterplot of the next subsection.

8.3.1 Results for the pooled organisations.

We start with a scatterplot of the error rate by the occurrence risk. Figure 8.1 shows the results.

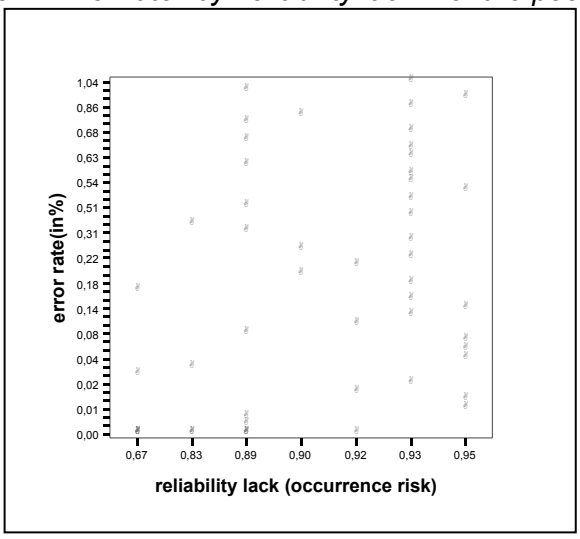
Figure 8.1: Error rate by 'reliability lack'** for the pooled organisations (n=52)*



* error rate on its original scale, **'reliability lack ón an ordinal scale

The scatterplot shows a relatively homogeneously distributed set of error rates. No explicit outliers can be seen, and no high concentration of error rates near 0 is seen. This can be checked in a scatterplot with the error rate also scaled as an ordinal variable.

Figure 8.2: Error rate by 'reliability lack'* for the pooled organisations (n=52)*



* error rate and 'reliability lack' on an ordinal scale

Figure 8.2 confirms what he saw in figure 8.1. Now, the error rate being relatively homogeneously distributed, both a Pearson and a Kendall correlation coefficient will make sense. Consistent with this observation, the pooled data showed similar values for the Pearson and the Kendall correlation between the error rate and the "reliability lack" (occurrence risk), as is shown in table 8.5 in the row "pooled (without: org 2)". Both correlations are significant at the 5% level, and have the plus-sign, as may be expected in a valid risk assessment. It is remarkable that the P-correlation is virtually zero when organisation 2 is included. Interpretation of this outcome suffers from the "reliability lack" (occurrence risk) for organisation 2 only being based on the internal control risk and an imputed value ("high") for the inherent risk.

Table 8.5: P- and K-correlations error rate x reliability lack (occurrence risk)

Organisation	K-correlation	p-value	P-correlation	p-value	n
1	-.066	.61	-.064	.71	37
2***	.078	.68	.016	.94	24
3	.100	.77	-.050	.91	8
4	-.246	.53	-.211	.65	7
pooled (without org 2)	.253*	.015	.326*	.018	52
pooled (with org 2)***	.232	.01	.003	.98	76

* significant at the .05 level *** For organisation 2 the correlations are only based on ICR; the value for IR has been imputed as "high".

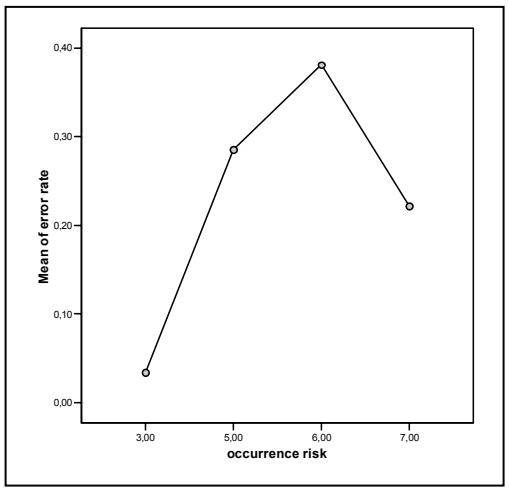
We will analyse the significant correlation for the pooled organisations, by comparing the means of the error rate per level of the occurrence risk, scaled according to the variable OR2, because in the "primitive" 'reliability lack' there are four categories with six or less cases. The results (organisation 2 excluded) are shown in table 8.6 and figure 8.3.

Table 8.6: Means of error rate by occurrence risk for 3 pooled organisations

occurrence risk (OR2)	Mean of error rate (%)	N	Std deviation
3	.034	6	.065
5	.28	19	.32
6	.38	19	.31
7	.22	8	.32
All levels	.282	52	.31

F-value: 2.2; p-value: .103, p-value for Levene's statistic for equal variances: .016:

Figure 8.3: Means of error rate by occurrence risk for the 3 pooled organisations



The analysis of the means shows that the positive correlation of table 8.5 does not imply a monotonous relation between the occurrence risk and the means of the error rate, but the positive relation is confirmed. An analysis of variance shows an F-value of 2.2 (p-value: .103), so the means do not differ at a significance level of 5%.

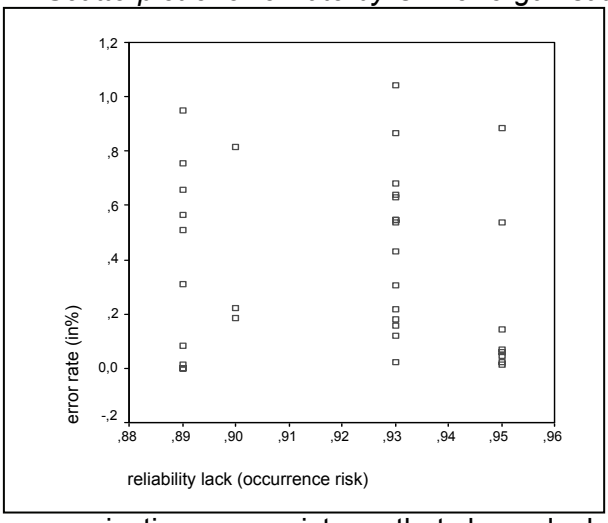
8.3.2 Results for the distinct organisations.

The results for the distinct organisations are given in table 8.5 above. All correlation coefficients, either Pearson or Kendall, are close to 0 and far from significant. This implies that there will be no significant differences with regard to the validity between the organisations. This is confirmed in testing the differences with Fishers z-transformation. The largest difference of correlations applies to organisations 3 and 4;

the p-value of the difference is .3. Also the difference between the correlations in organisation 1 and 4 has a p-value (.16) which makes it insignificant. So we cannot conclude that there is a difference in the correlations for the distinct organisations.

Just as we did for the pooled data, we plotted the “reliability lack” (occurrence risk) against the error rate in order to check the statistical validity of the Pearson correlation coefficient. The result for organisation 1 is given in figure 8.4. The (P or K) correlation (-.064 or -.066) of table 8.5 is fully confirmed by relatively homogeneous distributions of the error rate for the 4 available levels of the ‘reliability lack’; there are no obvious outliers.

Figure 8.4: Scatterplot of error rate by OR for organisation 1



The other organisations gave pictures that showed a lack of variability in the “reliability lack” (occurrence risk), which also can be seen from table 8.7. In every organisation, except for organisation 1, only two distinct levels for the "reliability lack" (occurrence risk) were found.

Table 8.7: Frequencies of the “reliability lack” (occurrence risk) per organisation

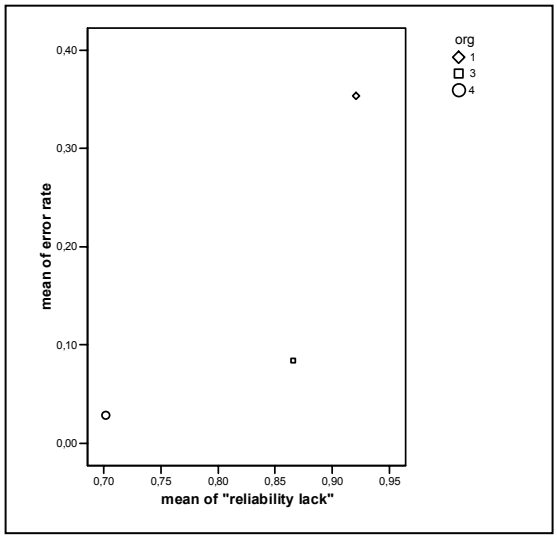
reliability lack	organisation 1	organisation 2	organisation 3	organisation 4
.67				6
.83		22	6	
.89	11			1
.90	3			
.92		2	4	
.93	16			
.95	9			
Total	39	24	10	7

This lack of variability implies that the relations found for the distinct organisation are not very informative for the full range of ‘reliability lack’, except for the data of organisation 1. Anyhow, per organisation there is no indication of a satisfactory correlation between occurrence risk and error rate.

Table 8.5 allows another striking observation: the correlations for the distinct organisations all are close to zero, except for those of organisation 4, but the correlations for the pooled organisations are significant in the expected direction. We will further analyse how this seemingly contradictory result can be explained, by calculating the correlations for the means of "reliability lack" and of the error rate per organisation. We did this for the 3 organisations of which both IR and ICR were known.

The resulting correlation is .799 (p-value: .411) , which gives an explanation for the resulting positive correlation for the pooled organisations of table 8.5. We further analyse this result, by plotting the means in the next figure

Figure 8.5:: Means of error rate by mean of "reliability lack" for 3 organisations.



We see that the correlation for the means per organisation is built on a monotonous relation of the means of the “reliability lack” and of the error rate per organisation. This indicates valid assessment of OR, but it lacks significance.

8.3.3 Controlling for materiality

In 2.3.1 and 6.3.3 we discussed the necessity of controlling for the materiality, when analyzing the correlation between error rate and occurrence risk. Therefore we categorized the materiality into 3 classes, with .98% and 1.2% as boundaries. With that categorization the middle class is very narrow, which is favourable for a raise in the correlation, assuming that controlling for materiality gives such an effect. It appears that in the highest class 24 cases out of 27 come from organization 2 for which, as known, the “reliability lack” could not be computed. Therefore we will restrict our analysis to two classes of materiality. Table 8.8 shows the results

Table 8.8: P-correlations reliability lack x error rate for classes of materiality

Level of materiality	reliability lack x error rate	size sub-sample
materiality< .98%	.427	18
.98%<=materiality<1.2%	.071	32
the two classes pooled	.306*	50

* significant at the .05 level

It appears that the correlation does not improve from controlling for the level of materiality. Just as in 6.3.3 we conclude that this is an indication that risk assessment is done with disregard of the level of materiality.

So in all, we have a significant correlation for the pooled organisations, but for the distinct organisations this correlation disappears. We conclude as follows.

Conclusion research question 2: validity with respect to error rate.
Validity with respect to the error rate is not strong.

Conclusion research question 5: increasing correlations when controlling for materiality
Contrary to the expectation, the correlation of error rate and reliability lack (occurrence risk) does not increase when controlling for materiality.

Conclusion research question 10: varying validity with respect to error rate.
Validity with respect to error rate does not vary over organisations; within the organisations all correlations between error rate and occurrence risk are insignificant.

Discussion
There is an indication of validity with respect to the error rate, when we look at the pooled data. But this indication is not convincing enough to justify varying audit effort with varying assessment of OR on this observation, because the relation found between error rate and reliability lack (occurrence risk) is not strong and not monotonous. Within organisations variation in audit effort due to varying assessments of the occurrence risk, is not justified by the results found. For 3 organisations a lack of variation in the 'reliability lack' may have hindered the investigation of validity, where for organisation 1 this variability was sufficient, but nevertheless only a small correlation was found.

8.4 Risk assessment and sampling risk

8.4.1 Definition of sampling risk

We define (categories of) the sampling risk as in section 6.4.1; table 8.9 helps recall the classification.

Table 8.9: Categories of sampling risk

category of sampling risk	boundaries
1: very low	SR < .05
2: low	.05 <= SR < .15
3: medium	.15 <= SR <.45
4: high	SR >=.45

We also recall the dependency of SR of the sample size and therefore of OR. But it appears that the correlation between these 2 variables is even negative (-.048); the correlation between SR and OR is positively affected (albeit to a negligeable extent).

8.4.2 Results for the pooled organisations.

We start our analysis by creating scatterplots (see figures 8.6) for the sampling risk and the standardised sampling risk against the "reliability lack" (occurrence risk),. The scatterplots shows a distribution of the sampling risks for all levels of 'reliability lack' (occurrence risk) which is of similar homogeneity as that found in 6.4.2. This implies that both the Pearson and the Kendall correlations make sense. The 4 correlations are given in table 8.10 (all analyses with organisation 2 excluded). The differences in n for the sampling risk (SR) and the standardised sampling risk (SR2) are caused by missing values for the size of the audit sample. So for instance in organisation 3 the size of the audit sample was missing for all cases, whereas in 8 cases an estimate of the error rate was available.

Figures 8.6: Sampling risk by "reliability lack" (occurrence risk), organisation 2 excluded

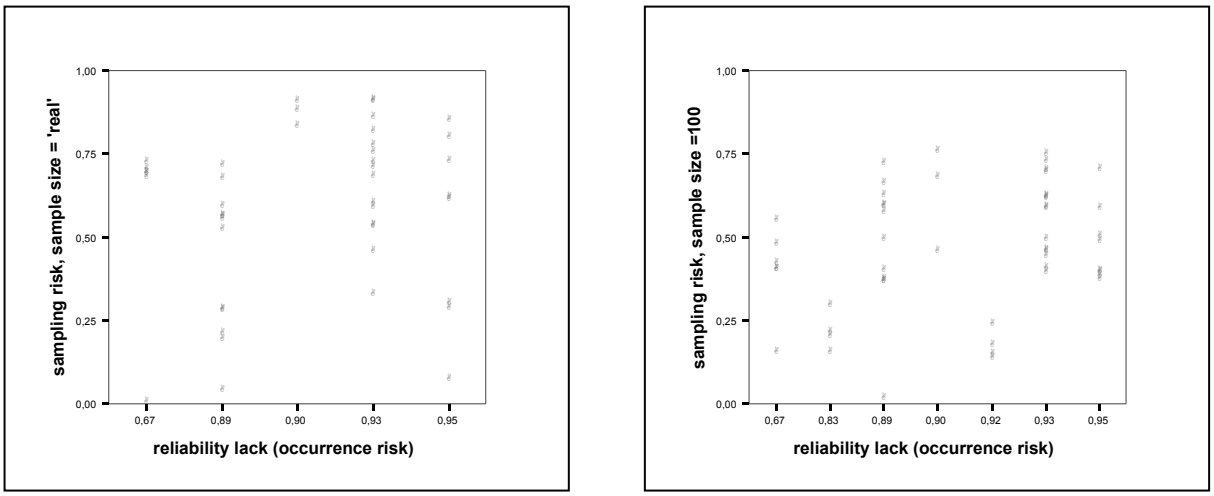


Table 8.10: Pearson and Kendall correlations of "reliability lack" (occurrence risk) by sampling risk (SR) and standardised sampling risk (SR2), organisation 2 excluded

Organisation	P-correlation SR	K-correlation SR	n	P-correlation SR2	K-correlation SR2	n
pooled	.047	.12	44	.23	.18	52

The correlations are positive, but none is significant; the smallest p-value is .08 for the K-correlation of .18.

We produced a cross tabulation of the categorised sampling risk with the occurrence risk (OR2). Table 8.11 gives the results.

Table 8.11: Sampling risk in categories by occurrence risk (ex organisation 2)

Sampling risk in categories→ Occurrence risk (OR2)↓	Very low	Low	Medium	High	Row total
3	1	0	0	5	6
5	1	0	4	10	15
6	0	0	1	14	15
7	0	1	2	5	8
Column total	2	1	7	34	44

The table shows only 7 cases on the diagonal, and relatively many (16) far from the same diagonal, so that the relation is weak or absent. This is also seen in the K-correlation, which has a value of only .018. The most striking element of the table is that 15 cases are far from the diagonal in the above-diagonal triangle, which means that 15 out of 44 cases must be seen as ineffective. Next to that, the categorisation of the 'reliability lack' and of the sampling risk has a slightly deflating effect on the correlation (it was .12 in table 8.10). This will also appear in the same analysis for the standardised sampling risk. Table 8.12 gives the results (organisation 2 excluded).

Table 8.12: Standardised categorised SR2 by occurrence risk (ex organisation 2)

Standardised sampling risk in categories→ Occurrence risk (OR2)↓	Very low	Low	Medium	High	Row total
3	0	0	4	2	6
5	1	0	8	10	19
6	0	2	5	12	19
7	0	0	4	4	8
Column total	1	2	21	28	52

The K-correlation decreases from .18 to .08 (p-value: .51). Again the number of (16) ineffective cases is remarkable. So with this categorisation of sampling risk and reliability lack, the relation between SR2 and OR2 is even weaker than in table 8.10. This especially with an eye on the remarkable number of ineffective cases, both for sampling risk and standardised sampling risk.

Conclusion research question 3: validity with respect to sampling risk
For the pooled organisations, assessment of the occurrence risk neither shows validity with respect to the sampling risk nor with respect to the standardised sampling risk,.

Conclusion research question 6: OR strongest relation with SR as a criterion?
A difference in the degree of validity of the assessment of OR with the error rate as a criterion or with the sampling risk as a criterion, cannot be shown.

8.4.3 Results for the distinct organisations.

We did the same analyses for the distinct organisations, as we did for their combination. So firstly we calculated all P(earson)- and K(endall)-correlations for the 4 organisations that were our units of analysis. Table 8.13 shows the results.

Table 8.13: Sampling risk (SR) x "reliability lack" and standardised sampling risk (SR2) x "reliability lack" by organisation

Organisation	P-correlation SR	K-correlation SR	n	P-correlation SR2	K-correlation SR2	n
1	.12	.14	37	-.10	-.076	37
2	-.19	-.14	16	-.14	-.030	24
3	.***	***	.	-.433	-.378	8
4	-.62	-.36	7	-.77*	-.54	7
pooled**	.047	.12	44	.23	.18	52

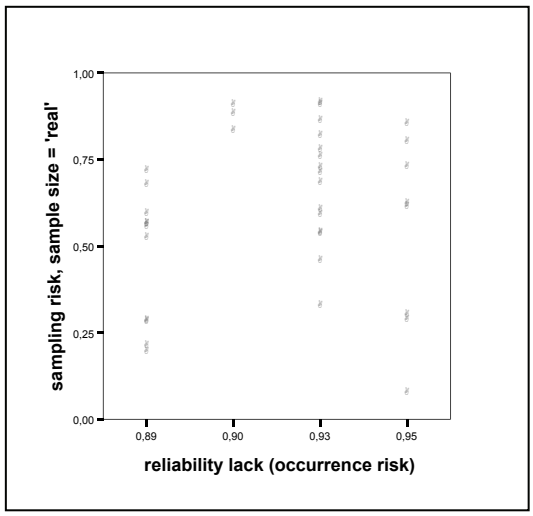
* significant at the .05 level (2-tailed) **organisation 2 excluded *** in this organisation we did not get the sample sizes

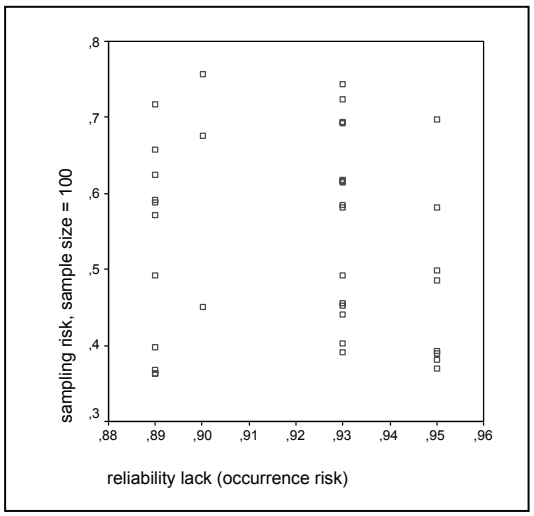
Only one correlation is significant, that for SR2 and "reliability lack" in organisation 4. But it has the unexpected minus-sign. This '-'-sign is to be observed in all cases, except for the two SR-correlations for organisation 1. This means a worsening of the degree of validity compared to that with the error rate as a criterion, where the correlations vary more around 0 (table 8.5). This worsening is parallel to that found in chapter 6.

In figures 8.7 we show the scatterplots for organisation 1, of the sampling risk and the standardised sampling risk against the "reliability lack" (occurrence risk). It indicates that the relation for organisation 1 deviates from linear. This is also shown by the eta, with the 'sampling risk' as the dependent variable and the 'reliability lack' as the independent variable; it is considerably larger than the P- or the K-correlation: eta=.39.

But at the same time the K- and the P-correlation are almost equal and both indicate that the correlation is negligible. So the relation is neither linear nor monotonous.

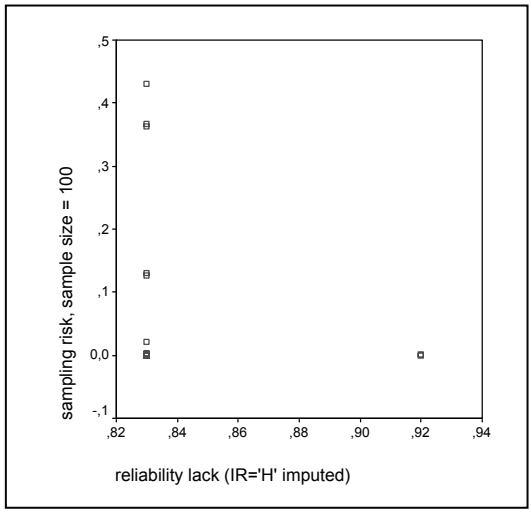
Figures 8.7: Sampling risk and standardised sampling risk against occurrence risk for organisation 1





The scatterplot for organisation 2 in figure 8.8 shows a far less interesting picture: only two distinct values for the 'reliability lack' and for the highest value only two cases (which both have SR=0).

Figure 8.8: : Sampling risk by occurrence risk for organisation 2



Organisations 3 and 4 show similar relatively uninformative configurations. Therefore, as in 8.3.2, we only continue the analysis for organisation 1. We do this by producing a cross tabulation: table 8.14.

Table 8.14: Occurrence risk by sampling risk in categories for organisation 1

sampling risk in categories→ Occurrence risk (OR2)↓	Low (.05< SR <=.15)	Medium (.15< SR <=.45)	High (.45< SR)	Row total
5	0	4	10	14
6	0	1	14	15
7	1	2	5	8
Column total	1	7	29	37

Again there are only a few cases on the diagonal and there are relatively many cases in the above-diagonal triangle, which indicates an ineffectiveness of the audits. As in the previous analyses, for the pooled organisations, we used the occurrence risk (OR2).

The cross tabulation shows that the highest categories for the sampling risk occur to a lesser extent at the higher occurrence risk, so there is a negative correlation. The K-correlation appears to be -.018 (p-value:.94); the P-correlation is -.11 (p-value: .68).

Conclusion research question 3: validity with respect to sampling risk
Neither for the pooled nor for organisation 1 risk assessment with respect to the sampling risk SR shows validity. For the other organisations the validity could not be established separately.

Conclusion research question 6: OR strongest relation with SR as a criterion?
The correlations between error rate and reliability lack virtually do not differ from those between sampling risk and reliability lack.

Conclusion research question 10: varying validity with respect to sampling risk.
Varying validity with respect to sampling risk over organisations can not be shown.

Discussion on validity with respect to sampling risk.
The correlation between sample size and OR being virtually 0, we expected on logical grounds, the validity with respect to sampling risk to be stronger than with respect to the error rate. This, because the sampling risk has the dimension of a risk, contrary to the error rate. This expectation is not satisfied: we again do not find such a difference, albeit that in this study the relevant correlations are approximately equal, where in the first study they differed (but in the unexpected direction).
On empirical grounds, as a consequence of the findings in the first study of our research, we expected that the correlations with the error rate would be stronger than the correlations with sampling risk. This expectation was not confirmed. As a consequence, no indication is found in this study, that risk assessment might be improved if the auditor would explicitly aim at predicting the error.

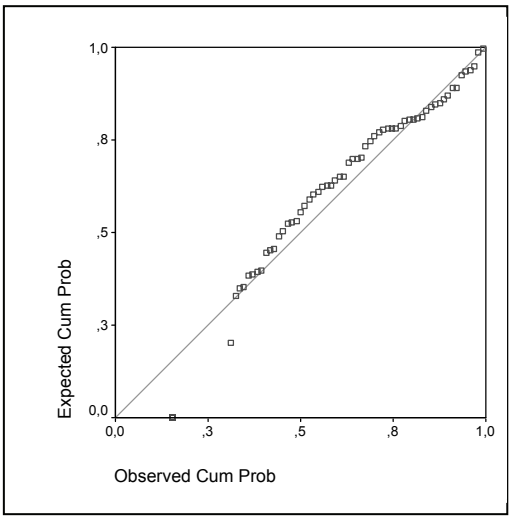
8.5 Risk assessment and conditional distribution of error rates

In chapter 2 we introduced the logic of using the distribution of the error rates, conditional on the assessed risk, as a validation criterion for risk assessment. Like in section 6.5, we show the results of this analysis only for the pooled organisations, because for the distinct organisations there are not enough data. We will give a scatterplot of the error rates for the four levels of the occurrence risk OR2 (table 8.2). We will also give the parameters of the beta distributions that can be fitted on the pooled data and on the data per level of assessed risk.

8.5.1 Results for the pooled organisations

We constructed a PP-plot (see footnote 12 in 6.5.1) of the distribution of the error rates, for a beta distribution, and calculated the best fitting parameters, both for the total of error rates and for the error rates per level of the occurrence risk. We show the PP-plot for the 4 pooled organizations and undivided for level of OR in figure 8.9.

Figure 8.9: PP-plot of the error rates for the beta distribution

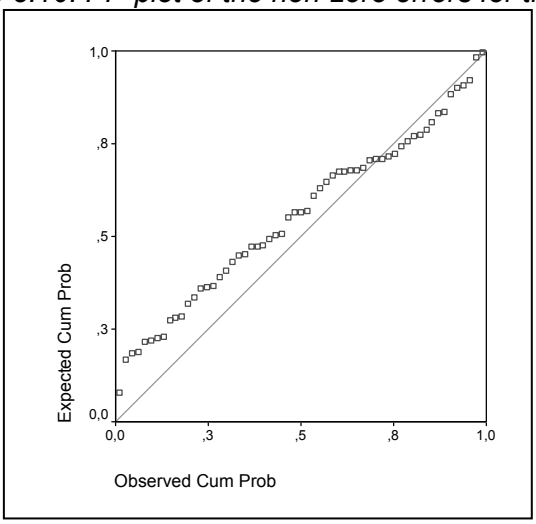


From figure 8.9 it can be seen that where some 20% of the observed error rates is smaller than the error rate corresponding to the first dot above the X-axis, in the fitted beta distribution values smaller than this observed value would have a much smaller probability. So here the fitted distribution and the actual distribution do not fit. But the fit improves and is good for the values that are larger than the 40 percent smallest observed error rates

As we discussed in 6.5.1, the poor fit for the lower cumulative probabilities is due to the 24% of the error rates equal to 0 (see table 8.15), where in a theoretical beta distribution none of any possible values has a probability larger than zero (due to its continuity). The parameters of the fitted distribution appear to be: a=.236; b=54,2. This means that the mean of the fitted beta distribution is equal to (a/a+b=).0043=.43%, which is exactly equal to the actual mean.

We repeat the analysis for the values of the error rate larger than 0. This gives figure 8.10 as a result. It shows less small errors than expected in the fitted distribution and more larger errors than expected
The parameters are: a=.379; b=60.3

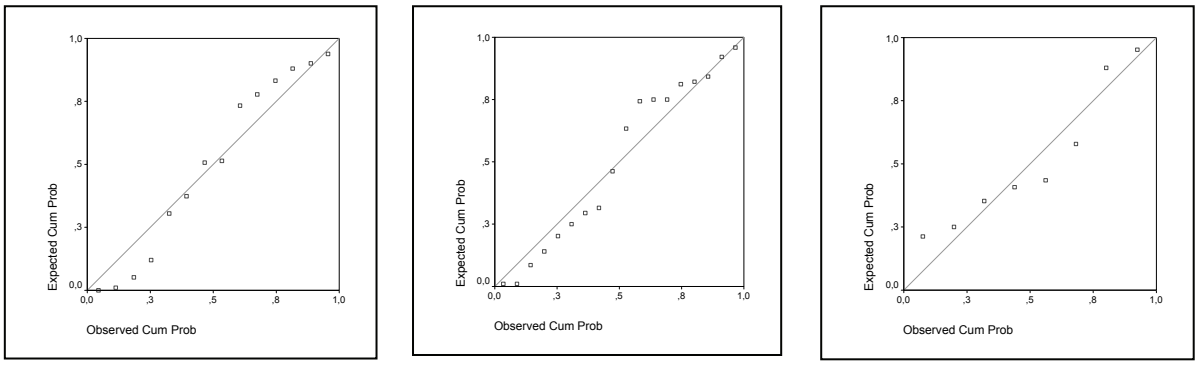
Figure 8.10: PP-plot of the non-zero errors for the beta distribution



We applied the same analysis for three levels of the occurrence risk: OR2= 5, 6, 7; OR2 as defined in 8.3.1 . As there were only two cases with a non-zero error rate for

OR2= 3, we left out this fourth level from analysis (no other levels for OR2 were observed). The analyses again apply to the non-zero error rates. Because in this analysis the relation between distribution of error rate and assessed occurrence risk is object of investigation, we leave out the data of organisation 2 (as before, because IR is missing) In table 8.15 we have also left out organisation 2.

Figures 8.11: P-P plots for levels of the occurrence risk 5, 6 and 7, organizations 1,3,4



OR2=5, n=14, a=1.5, b=374 OR2=6, n=18, a=1.8, b=444 OR2=7, n=8, a=.48, b=218

The figures 8.11 again show considerable differences between the distributions of the error rates for the three levels of the occurrence risk that were observed. We make these differences more visible in table 8.15, by presenting the means of the non-zero errors and by including the zeros in the analysis, so that we can combine the mean of the beta-distribution for the non-zeros with that for the zeros..

Table 8.15: Relative rates of 0 and beta distributions for the non-zero errors by occurrence risk (OR2) for organisations 1,3,4.

Occurrence risk→ (Beta-) distribution↓	3	5	6	7	All*
#(greater than 0)	2	14	18	8	42
a-parameter		1.457	1.79	.484	1.271
b-parameter		374.5	443.6	217.7	363.5
mean non zeros		.00389	.00403	.00222	.00350
#(equal to 0)	4	5	1	0	10
rate of 0 (%)	67	26	5	0	24
weighted mean		.00287	.00382	.00222	.00283

*all cases where OR2, the occurrence risk was known

The outcomes are consistent with table 8.6. Firstly the weighted means in the last row in table 8.15 are equal to those in table 8.6. Secondly, also for the nonzero errors, the means show a peak at an occurrence risk of 6.

Further we can see that the rate of zeros is monotonously decreasing with an increasing occurrence risk. Together with the non-monotony of the relation between occurrence risk and error rate, this may be seen as an indication that risk assessment has more predictive power for the existence of an error than for its size . The same conclusion has been drawn in 6.5.1 (table 6.18).

The high sum of a and b parameters in all columns of table 8.15, compared to that sum for the fitting distribution for all errors is due to the leaving out of organisation 2. In this organisation an error of 6.25% occurred (as the maximum) whereas in the other organisations the maxima are 1.04%, .31% and .16% respectively. With such small maxima the b-parameter should be large, to get sufficient weight for the small errors.

8.5.2 Results for the distinct organisations

There were not enough cases to perform this analysis on the level of the distinct organisations. It asks for an analysis at the 4 distinct levels of OR2, for 4 units of analysis. So on the average, in the optimistic case that we would include organisation 2, we would only have 76/16=5 cases to base the distributional analysis on. In the more realistic case of excluding organisation 2, we would have to base our analysis on the average on only 52/12=4 cases per level, organisation. Therefore, it makes no sense doing this analysis per organisation.

We come to a conclusion regarding this criterion.

Conclusion research question 4: validity with respect to conditional distribution
In the second study risk assessment shows a satisfactory validity with respect to the prediction of the occurrence of errors, less with respect to their size.

Discussion on validity with respect to the conditional distribution of the errors
The indication we discussed in 6.5.2, that an auditor's risk assessment might be better in predicting the size of an error than in assessing a risk is not confirmed by the findings of this chapter. Validity with respect to error rate, with respect to the sampling risk and with respect to the conditional distribution of the errors all are of similar quality, where in chapter 6 the validity with respect to error rate was considerably better than with respect to sampling risk.

8.6 Moderator variables?

Only for organisation 1 data on the possible moderator variables were sufficiently available the number of days spent for the audit and the type of transaction. In the other organisations these were only available in numbers that were too small to allow statistical analysis. This is also due to the fact that combinations of 3 variables are necessary: the moderator and the two correlated variables. So we will only analyse possible moderators for organisation 1.

8.6.1 The influence of the effort for organisation 1

We made a dichotomy with respect to 'effort': the effort was "small" when at most 12 days were spent for the risk analysis, and "large" when more than 12 days were spent. With this classification we found the results given in table 8.16.

Table 8.16 P- correlations of "reliability lack" (occurrence risk) with 3 variables for levels of "effort"

"reliability lack" with→ for↓	error rate	sampling risk (real n)	sampling risk (n=100)
"effort" <= 12 days	.109 (n=13)	-.011 (n=13)	.095 (n=13)
"effort" >12 days	.110 (n=9)	.196 (n=9)	.123 (n=9)

Evidently the correlations differ only marginally, which leads to the conclusion that "effort" is not a moderator.

Conclusion research question 8: is effort a moderator?
The variable is not acting as a moderator for the validity of the assessment of OR with respect to the criteria error rate and sampling risk.

Discussion:
"Effort" may be correlated to "size". When this is the case, more effort might be "compensated" by larger size (complexity), causing a net result of no moderator effect. But the actual correlation of "size" and "effort" (in organisation 1) is -.013[19].. So this possible explanation does not hold. We must stick to the conclusion that (in organisation 1) "effort" is not a moderator.

8.6.2 The influence of type of transaction for organisation 1

As explained in 8.1.1 (research question 18), our data concern two types of transactions: obligations and payments. We subdivided the data after this quality and compared the relevant correlations. The result is given in table 8.17.

Table 8.17 P- correlations of "reliability lack" (occurrence risk) with 3 variables for two types of transactions in organisation 1

"reliability lack" with→ for↓	error rate	sampling risk (real n) SR	sampling risk (n=100) SR2
obligations (n=15)	-.213	-.364	-.441
payments (n=22)	.067	.062	.064
p-value difference	.224	.116	.077

The differences for obligations and payments in the correlations all are considerable, but none has a p-value less than 5% (the smallest, that for SR2, is 7.7%). So we cannot conclude that 'type' is a moderator variable.

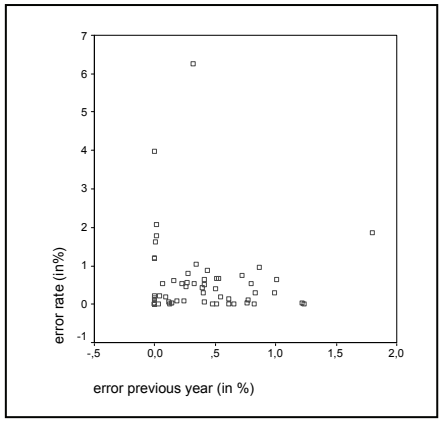
Conclusion research question 18: type of transactions a moderator?
We cannot conclude that type is a moderator variable for organisation 1.

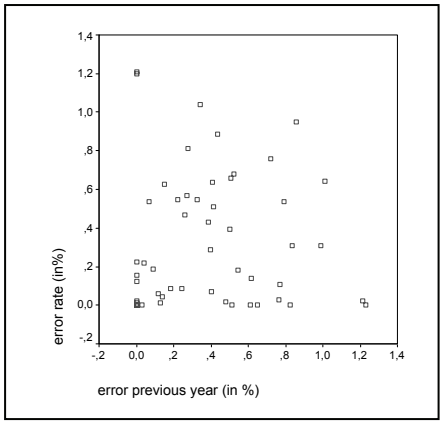
8.7 Error previous year a predictor for error this year?

In 8.1.1 we introduced research question 19 on the possibility that the error 'of the previous year' is a predictor for the error 'of this year', to investigate a possibility of having another predictor than risk assessment for the for the error of the account to be audited (the error 'of this year'). In our data collection we had access to the error of the previous year in the organisations 1 and 2. So for these two organisations we can investigate the relevant relation. We start with giving the relevant scatterplots: for the pooled and for the distinct organisations

[19] This actually confirms the logic of our choice not to use 'size' as a moderator; see 5.2.3.

Figures 8.12: Scatterplot error against error previous year for pooled organisations (1 and 2) (n=68)

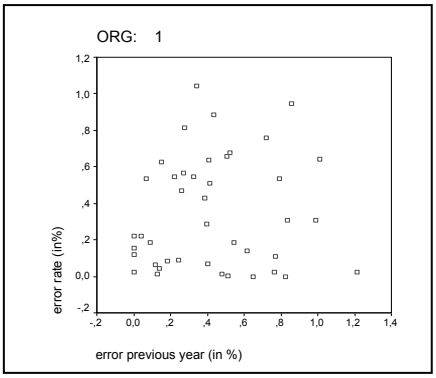


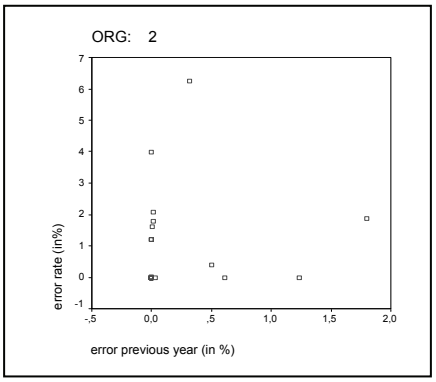


P-correlation=.026 (p-value= .83, n=68) outliers omitted: P-correlation=.15, (p-value=.24, n=62)

The correlation corresponding to the first scatterplot is .026 (n=68). When we leave out the error rates (of this year) larger than 1.6%, the resulting correlation is .15 (n=62). It is interesting to see how the two distinct organisations contribute to this correlation. We show the scatterplots in figures 8.13.

Figures 8.13: Scatterplot error against error previous year for organisations 1 and 2





P-correlation=.089 (p-value= .58, n=42) P-correlation=.128 (p-value=.53, n=26)

Maybe the most striking of the patterns in the scatterplot of figures 8.13 is given by that of organisation 2: where last year the errors could be (almost) any size, this year the error is (almost for all cases) close to zero; where this year the error can be any size, last year the error was very close to zero. The scatterplot for organisation 1 shows an almost completely random distribution of the pairs of errors over the graph.

Conclusion research question 19: error of the previous year predicts error this year?
The error of the previous year cannot be used as a predictor for the error of this year.

Discussion
The findings with respect to the predictive qualities of the error of the previous year are very revealing. They show that in addition to the fact that the validity of risk assessment is questionable, the error of last year does not provide refuge from this finding. This observation can easily be interpreted as a consequence of the existence of two effects:

- the existence of the negative regression effect over years as we discussed in 8.1.1 research question 19: accounts in some year with high (low) error tend to have low (high) error next year, due to improvement (deterioration) of the administrative processes;

- accounts with a stable quality of the administrative processes also (might) tend to be stable in error rate.

We believe the interpretations are worth to be considered more deeply. We will do that in chapter 9 (the negative regression) and in our final chapter, 10.

8.8 Summary, Discussion and conclusions

8.8.1 Summary and discussion

The aim of this chapter was, to report the findings of a replication of the study of chapter 6. The conclusions made per research question are not very encouraging.

The relation with 'audit position' gives only very weak correlations. This could be caused by the data being not very strong, because only a few cases with 'audit position' 'not OK' occurred.

Only for the pooled organisations a significant correlation is found between the error rate and the occurrence risk. The practical value of this is to be questioned, because all correlations within organisations are close to zero. Again the risk assessment seems to be performed with disregard of materiality, and the relation between occurrence risk and error rate is not monotonous.

The relation between assessed risk and sampling risk is unsatisfactory, both for the pooled and for the distinct organisations. Based on the sampling risk, 15 out of 44 cases had to be labelled as 'ineffective', based on the standardised sampling risk this number was 16 out of 52. Within the distinct organisations the relation between occurrence risk and sampling risk is absent, as is the relation with the error rate. Here again there was an alarming number of 'ineffective' cases.

The analysis of the conditional distributions shows a modest validity for the pooled organisations. This analysis was not done for the distinct organisations, due to an insufficient number of cases.

The potential refuge of the error of the previous year appears not to work, probably due to negative regression effects. The analyses reveal another trait of the field in which an auditor has to work: he has to assess risk in a situation where the error rate only varies in a very limited range. In the past he had to do with many error percentages well below 1% and in many cases he will have the experience that they will probably be in the same domain again in the year his audits apply to. Practice of risk assessment aims at gaining more information on the error rate, so the auditor is confronted with the question whether error rates within such a small range can be predicted more precisely than that they probably will be below 1%. Essentially this is what the auditor aims at. And this necessity of precision is also implied by the audit methodology, which takes the absence of material errors as a point of departure (see also the 6th consideration in section 6.7). In our chapter 10 we will go further into the rationality of goals implying such a precision

8.8.2 Conclusion

The replication gives a weaker conclusion than in chapter 6, as to the validity of risk assessment: for the pooled organisations there is an indication of validity, but within the organisations validity could not be established. Remember that this validity was found in chapter 6 (albeit with quite some caveats). The conclusion of chapter 6 is confirmed

and inescapable, that an organisation cannot rely on its risk assessment as a justification for a decrease in substantive audit, unless it has strong empirical evidence on its validity.

8.9 Summary of findings

The next table gives an overview of the findings reported in this chapter, organised by research question.

Table 8.18 Overview of the findings in chapter 8

Research question	Result
1: To which degree will risk assessment (OR) correlate with the 'audit position' of the error in a sample?	Assessment of OR only shows insignificant correlations with the variable 'audit position'; it is not valid with respect to this variable (section 8.2).
2: To which degree will risk assessment correlate with the error rate in the audited account?	Risk assessment and error rate do not correlate very strongly; lack of monotony in the relation may cause problems (section 8.3).
3: To which degree will risk assessment (OR) correlate with the sampling risk (SR)?	Neither for the pooled organisations, nor for organisation 1 the assessment of the occurrence risk is valid with respect to the sampling risk (section 8.4).
4: To which degree will the distribution of the error rate vary with the level of assessed risk?	The distribution of the error rate, conditional on the assessed level of risk, varies with risk assessment in the expected direction (section 8.5).
5: To which degree will the correlation between error rate and risk assessment increase when calculated for groups of accounts with the same level of materiality compared to the correlation for the whole group of accounts?	Contrary to the expectation, the correlation of error rate and OR does not increase when controlled for levels of materiality. (8.3.3)
6: To which degree will the occurrence risk (OR) show a relation with the sampling risk, which is stronger than that between OR and the 'audit position', or OR and the error rate?	Both for the pooled organisations and for the distinct organisations the answer to this question is negative (8.4.3)
8: To which degree will the level of validation increase with the effort put in the assessment of OR?	The validity of risk assessment does not improve with increasing effort spent in this activity. (8.6.2)
10: To which degree will the level of validation vary over organizations?	This question could not be answered for the 'audit position' Both with respect to error rate and with respect to sampling risk a variation in validity over organisations could not be established, but only for one organisation these validities could be established in a satisfactory way(8.3.2, 8.4.3)
18: Is type of transactions a moderator?	We cannot conclude that type is a moderator variable for organisation 1. (8.6.3)
19: Is "error of the previous year" a predictor for "error this year"?	The error of the previous year cannot be used as a predictor for the error of this year. (section 8.7)

Chapter 9: The Predictive Power of System Tests.

Four relatively complete actions in three studies were taken in our research into the 'real risk': (1) validation of the classical assessment of OR, (2) an attempt to improve classical assessment of OR by decomposition of the assessment over risk indicators (1st and 2nd action formed the first study), (3) a replication of the validation of classical assessment of OR in the second study and (4) investigation of the predictive power of system tests for the error rate in the third study. Varying validity of classical risk assessment per organisation (1st action) and problematic results with risk indicators (2nd action) led to the choice for the replication (3rd action).
In this chapter we report the results of the 4th action. It does not directly aim at investigating risk assessment itself, but at investigating a standard method for underpinning risk assessment.

9.1 Introduction

In 5.2.5 we discussed system tests in relation to risk assessment. We concluded that system tests should be predictive for the error rate, for to be fit as an underpinning of risk assessment. After the chapters 6 and 8, with their positive and negative results with respect to the validity, the question is more urgent whether system testing really is fit for this underpinning.

Therefore we try to answer the question whether system testing has predictive qualities for the error rate. If the answer is positive, we have found a way to improve on the validity of risk assessment. Because, even if in all cases (where relevant, because of an assessment of 'low' or 'medium') system testing was used as underpinning, it could be taken more advantage of by giving it more weight than usual. And obviously, in cases where system testing was not included in risk assessment, the possible improvement speaks for itself.

In this chapter we will investigate these predictive qualities. For that purpose 37121 records with data of dual purpose tests were available, by the courtesy of a Dutch governmental audit organisation. We analysed these data with respect to the predictive value of system errors for substantive errors. As in the first two studies of this research, we agreed on maintaining anonymity.

The chapter is organised as follows.
In section 9.2: some definitions, an overview of the data, a discussion of generalisability and a subdivision of research question 17 are given,
in section 9.3 the predictive qualities at the level of a transaction are reported,
in section 9.4 the predictive qualities at the level of the account are given and
in section 9.5 conclusions are drawn.

9.2 Definitions and data

The logical level to analyse the relation between system errors and substantive errors, is a transaction. But we will also aggregate the system errors and corresponding substantive errors to the level of the account to which they belong and analyse these

aggregates. Therefore we will define "system errors" and "substantive errors" both at the level of a transaction and at the level of an account.

9.2.1 Definitions

Definition of dual purpose test:
A test on an individual transaction in which both the operation of the controls (the administrative system) is checked and the correctness of the *book-value* is investigated.

Definition of system error
A transaction contains a system error when it has not been handled in accordance to the rules of the administrative organisation.
"System error" and "compliance error" are synonyms in this definition. In 9.2.2 we give the categories that were used and given in the data that were available to us. We did not have a more detailed specification.

Definition of fraction of system errors
The fraction of system errors in an account is the number of transactions with a system error divided by the total number of transactions in the account.

Definition of substantive error
A transaction contains a substantive error when at least one of the following deficiencies occurs:

- (part of) the transaction was not in accordance with the demands of legality and/or regularity
- the transaction was settled for an incorrect amount
- the transaction is reported with an incorrect amount.

In 9.2.2 we give the categories that were used and given in the data that were available to us. Again, we did not have a more detailed specification.

Definition of taint
A taint is the size of a substantive error divided by the size for which the transaction has been booked; in other words: the fraction of the book-value for which the transaction is in error.

Definition of fraction of substantive errors
The fraction of substantive errors in an account is the number of transactions with a substantive error divided by the total number of transactions in the account.

Definition of mean taint
The mean taint is the mean of all taints in the sample: the sum of these taints divided by the number of records in the sample[20].

Remark

In a dual purpose test the operation of the administrative processes and their outcome are tested more or less simultaneously. This leaves unimpeded that in the process the system part precedes the outcome and that it can be treated independently from the outcome (as is done in system testing), just as the outcome can be treated independently from the system part (as is done in substantive testing). This ‘independence’ is used when we try to explain the substantive error of year t from the system error of year t and the substantive error of year t-1 (see 9.4.3).

[20] With this definition the mean taint found in a sample is an unbiased estimator of the error rate in the population, provided the sample has been drawn as a monetary unit sample.

The next tables give an overview of the definitions

system errors

level	occurrence	size
transaction	yes/no	not
account	fraction of records with system errors	applicable

substantive errors

level	occurrence	size
transaction	yes/no	taint
account	fraction of records with substantive error	mean taint

9.2.2 The data

From one of the participating governmental organisations, mentioned in chapter 1, we got data of dual purpose tests performed on a monetary unit sample from the accounts of five years: 1995 up to 2000. A record in these dual purpose tests contains data with respect to compliance to the regulations and, if a substantive error occurs, its size.

The compliance errors were given in the following categories, (translated from the Dutch; no extra information on the errors was available):

- authorisation not sufficient
- not complied to accounting regulations
- incomplete accounting files
- essential data are missing
- insufficient quality of offering procedures
- miscellaneous procedural errors
- appropriation/payment too late
- certificate not filled out
- errors in administration of obligations

One of the problems with these categories was that they are not exclusive. Therefore answering a question like: "Can system errors be ordered as to their risk of causing a substantive error?" was impossible. In spite of this shortcoming, we used the data as we got them, simply because these categories were the ones in use in the audit department. And we took the limitations on our analyses for granted, limitations like that we had to confine ourselves to the existence of a system error and that we could only get some ordering of the seriousness of this error by using the number of system errors reported for a transaction. Also, seriousness because of the nature of a system error could not be established, because of the overlap of definitions. As to the substantive errors, data on the size of the transaction and the size of the substantive error (if there) were available.

The substantive errors were given in the following categories, (translated from the Dutch; no extra information on the errors was available):

- undue charge of budget
- appropriation or payment too large due to false application of regulation
- appropriation or payment too large due to double obligation or payment
- appropriation or payment too large due to untimely processing of new data
- wrong size of appropriation or payment due to other causes
- other financial errors
- appropriation or payment too small.

Table 9.1 gives an impression of the frequencies and fractions of system errors and substantive errors in the transactions for the years 1995 up to 2000.

Table 9.1: Frequencies of substantive errors and system errors per year

Year	Number of substantive errors	Percentage of substantive errors	Number of system errors	Percentage of system errors	Number of records
1995	253	1.8	2949	21.5	13743
1996	64	1.2	1069	19.3	5531
1997	31	0.5	559	9.6	5818
1998	97	2.0	562	11.5	4882
1999	36	0.5	530	7.4	7147
Total	481	1.3	5669	15.3	37121

It appears that in every year many system errors occur, where at the same time considerably less substantive errors are found. The audit opinion is not dependent on the number of errors, but on the size of the error. We give an overview of these sizes (expressed as mean taints) at the level of the accounts in table 9.2.

Table 9.2: Mean taints per account, per year

Year	-0.5% <mean taint<=0	0	0<mean taint<=0.5%	0.5%<mean taint<=1%	mean taint>1%	Number of accounts
1995	5	23	15	2	4	49
1996	1	26	9	1	6	43
1997	3	33	4	2	5	47
1998	1	37	3	1	8	50
1999	1	31	6	1	5	44
Total	11	157	32	7	26	233

Table 9.2 shows that the 37121 records were grouped in 233 accounts. Relatively many had a mean taint larger than 1%. Not all of these accounts were object for audit opinion, but still the 1% level for a mean taint is interesting, because as a rule for larger accounts 1% is the level of materiality in governmental audits. So for this study the question is interesting whether the quality of the operation of the system is a predictor for exceeding this 1%-level.

9.2.3 Generalisability

In this study the definitions as used by the governmental audit organisation are our point of departure. Even if there were some problems with these definitions, like the overlap in categories of system errors, we took them for granted because of the content validity (see Babbie, 1995). If we had tried to solve the ambiguities by changing definitions, we would have taken a greater distance from the actual audit practice of the audit organisation.

Again, like the previous ones, this study is based on a "convenience sample": we took the opportunity of being provided with a host of data in which an answer to an interesting question was hidden. This question: “Do system tests have predictive power for substantive errors?” in principle applies to organisations and accounts/transactions produced by the organisation. Our data only apply to one audit organisation, only selected by the ‘convenience of the availability of data’. The data form a MUS-sample from the relevant accounts. So there are strong limitations on generalisability to organisations, which we took, as an inevitable consequence, for granted. We could and

did not aim at external validity (see Cook and Campbell 1979). But the generalisability from sample (from an account) to the total account is guaranteed by considerable sample sizes and appropriate sampling.

Again, like in the first two rounds, this way of "selection" of data does not mean that the findings of this study will not be applicable to a wider population of audit organisations and the way they perform system testing. This population is characterised in principle by the same way of acting in auditing as is prevalent in the audit organisation of this study; but again it is clear that the range of this generalisation cannot be determined.

9.2.4 Research questions

In chapter 5 we formulated research question 17: *"To what extent are system tests predictive for the error rate and valid in that sense?"* In this part of the study we can give a specification of this question and extend it in the following two (sub) questions:
17.1 Is the absence of a system error in a transaction a guarantee for the absence of a substantive error in the same transaction?
17.2 Is the quality of the operation of the administrative system at the account level predictive for the fraction of substantive errors and/or mean taint?

We include two extra research questions: the first of these was already introduced in the previous chapter as r.q. 19. Because we can extend this question to the fraction of errors, we give it the number 20. The second of these represents candidates for the most logical predictors of the substantive error. A positive answer to both questions would falsify the negative regression we discussed in section 8.7(a.o.).
Research question 20 Are the substantive errors (fraction, mean taint) of the year t-1 a predictor for the substantive errors (fraction, mean taint) of the year t?
Research question 21 Does the prediction of the mean taint (t) improve by using both mean taint (t-1) and fraction of system errors (t)

9.3 Predictability at the transaction level

We analysed the predictability of a substantive error from a system error by looking at the rate of correct predictions and at odds ratios at the transaction level. Due to the definitions, it cannot be excluded that one error can both be seen as a system error and a substantive error. When this occurs, we call it a 'self evident combination'. Self evident combinations inflate the predictability we are after. We will discuss this in 9.3.3.

9.3.1 Correct predictions

We say that a "prediction" is "correct" when one (or more) system errors in a transaction coincide with the occurrence of a substantive error, or when the absence of a system error coincides with the absence of a substantive error. Table 9.3 shows how well system tests predict in the year 1999.

Table 9.3: Substantive error predicted by system error

1999	no system error	system error	Total
no substantive error	6600 (99.7%)	511 (96.4%)	7111 (99.3%)
substantive error	17 (0.3%)	19 (3.6%)	36 (0.5%)
Total	6617 (100 %)	530 (100%)	7147 (100%)

In 3.6% of the 530 cases where a system error was found, also a substantive error occurred; so in these cases "the error was predicted" (by the system error). Also for 0.3% of the 6617 cases where no system error was found, still a substantive error was found. In these cases "the error was missed" (due to the absence of a system error). The percentages of "errors missed" and "errors predicted", defined in this sense, are given in table 9.4 for the five years covered by our data. So table 9.4 gives the percentages in the row “substantive error”, from all the tables that are similar to table 9.3. All relations are highly significant: p-value=0 for the corresponding Fisher’s exact test. Evidently this high significance is also caused by the large sample sizes.

Table 9.4 Substantive error predicted by system error for five years

Year	error predicted %	error missed %	all errors %
1995	4.0	1.3	1.8
1996	3.4	0.6	1.2
1997	2.0	0.4	0.5
1998	7.3	1.3	2.0
1999	3.6	0.3	0.5
mean	4.0	0.8	1.3

It appears that the rate of correct predictions of a substantive error is equal to 4%, averaged over the five years. So in 96% of the cases where a system error was found, no substantive error occurred. On the other hand, only in 0.8% of the cases where no system error occurred, a substantive error was found,

Looking at table 9.4 another way shows that, given that a system error is observed, the probability that a substantive error will be found is 5 times larger (on the average) than when no system error is found. The column ‘all errors’ gives the probability of the occurrence of a substantive error if the information of system tests would not be used. This probability is more than 3 times smaller (on the average) than that of the cases with ‘error predicted’. But the probability of missing an error is still about 60% (on the average) of the same probability when no system test is used.

9.3.2 Odds ratios

In the analysis of 9.3.1 we concentrated on the predictive power in terms of the improvement of the rate of correct predictions, when the information on the outcome of the system test would be used. Another way of analysing is by means of the so-called "odds ratios". When a person says: "10 against 1 that it will rain in the coming hour" he is also stating something about the probability of rain to come, not as a probability, but as a ratio of probabilities also called the "odds". So odds ratio is a ratio of ratios.

In our research we will analyse these odds ratios by looking at the odds of a substantive error, when a system error was observed compared to the odds of a substantive error, when no system error was observed. In table 9.3 the odds for a substantive error are: 19/511 when a system error is found and 17/6600 when no system error is found. The odds ratio is: 19/511:17/6600 = 6600*19/17*511=14.435.
The ratio of these odds of course is informative for the effect of the system error: the odds may be expected to increase when a system error is observed. In our analyses the odds for a substantive error when a system error was observed are in the numerator of the ratio. So then the effect of a system error on the existence of a substantive error is mirrored in an odds ratio larger than 1. Table 9.5 gives the odds ratios for the five years and for the pooled data.

Table 9.5 odds ratios for substantive errors for five years

year	Odds ratio
1995	3.291
1996	5.519
1997	5.258
1998	5.992
1999	14.435
All years	4.361

When a system error is present, the odds for the existence of a substantive error is more than four times larger than when no system error is present. This result, for the pooled years, can be seen as the relatively strong relationship. The p-value of this odds ratio is 0, so the outcome is highly significant. Especially the odds ratio for 1999 is very high; it is a kind of climax in an improving tendency for the odds ratios over the five years.

It is possible to test the difference for the five years for the odds ratios, by means of the Mantel-Haenszel statistic (MHS, see Bishop et al p.147). This MHS appears to be 298, its corresponding p-value is 0. So we can reject the null-hypothesis that the odds ratios are homogeneous over the years: it confirms our observation of the improvement in predictive power for the occurrence of substantive errors over the years.

9.3.3 Self-evident combinations

A "self-evident combination" occurs when an error can both be seen as a system error and as a substantive error. For instance: the error "disposal/ payment too late" can occur in one transaction both as a system error (with the same label) and as a substantive error: "inappropriate application of regulations" (this occurred 20 times). Seen from the viewpoint of predictability, this may imply an increase in predictability, because when this system error is found, it is 100% sure that the corresponding substantive error occurs in the same transaction. But with other system errors, for instance a shift of a booking from the appropriate sub-account to an inappropriate one, it depends on the level at which the account is audited: when the two sub-accounts are part of the account under audit, there will be a system error, but no substantive error. So we cannot say that the predictability for self evident combinations is 100%. We could not establish to which extent the predictability is affected, so we only can note that it will be affected, but to an unknown extent.. As self-evident combinations do not mean that the relations found in the previous two sub-sections are not fit for validation of risk assessment by system tests, we take this inflation of the predictability for granted.

9.3.4 Predictability of the error size

There is a major difference between the occurrence of an error and its size. The audit opinion is not based on the occurrence of errors but on the size of the total error. So far our analyses concerned the occurrence of errors; in this subsection we will pay attention to the size of the substantive error and its dependence on the occurrence of a system error. We will do this by comparing the mean taint in those cases (records) with a substantive error where also a system error occurred, with the mean taint in those cases (with substantive error) without a system error. It may be expected that correct operation of the system will mitigate a substantive error, even if it could not prevent it. (Note that it does not make much sense to compare the mean taint in all cases where no system error occurred with the same mean in all cases with a system error, because there are so many cases without a system error that the mean taint of these cases will

always be much smaller than the mean taint in the cases where a system error occurred.)

By restricting ourselves to the cases where a substantive error occurred, we actually look at whether correct handling of the administrative procedures mitigates the substantive error, compared to those cases where the administrative procedures were violated. (Note also that this "search for mitigation" is not conclusive for the error in the total account, because then the occurrence of substantive errors also plays a role: many records with small substantive errors can aggregate into a material error. Only when we do the analysis at the account level, we can and will incorporate this (see section 9.4).)

Table 9.6 gives the mean taints for cases with a (or more than one) system error against the mean taint for cases without a system error. We took the error in its absolute (irrespective of the sign) and in its original (the sign included, so also possibly negative) value, because both can be in use in auditing.

Table 9.6: Mean taints against occurrence of system errors

Year	system error	mean abs taint (%)	n	F	p-value	mean taint (%)	F	p-value
1995	No	32.4	135	5.2	0.024	26.3	1.3	.256
	Yes	43.8	118			32.9		
1996	No	16.1	28	1.7	0.20	14.9	1.9	.175
	Yes	26.9	36			26.5		
1997	No	47.6	20	.95	.338	29.7	.98	.331
	Yes	62.7	11			50.6		
1998	No	36.3	56	2.9	.093	33.2	3.2	.074
	Yes	50.5	41			49.2		
1999	No	40,9	17	.57	.455	40.0	.65	.426
	Yes	53.1	19			53.1		

In all years there is a tendency for the mean taint in the presence of a system error to be larger, but only one difference is large enough to reach the 5% significance level: that for the absolute taints in 1995. The sign test gives a p-value of 3.13 percent for 5 favourable outcomes out of 5. So the number of differences in expected direction is significant. Except for the year 1995, it hardly makes difference whether the absolute or the original value of the taints is used.

9.3.5 Conclusions on predictability at transaction level

The rates of correct predictions in table 9.4 and the odds ratios in table 9.5 show a clear relation between the occurrence of a system error and the occurrence of a substantive error. Moreover the odds ratio shows a clear "learning effect" over the years. Of course this effect was also present in table 9.4.

Some mitigating influence on the size of the substantive error from correct application of administrative procedures, could be shown. With an eye on the sometimes high percentages with which substantive errors occur, in all years (see table 9.4), such a mitigating influence is of importance. Table 9.6 shows that there is no "learning effect" over years for the size of an error, at the level of a transaction.

Conclusion research question 17.1: absence system error guarantee absence substantive error?

The answer to this question can be moderately positive:

- On one hand we did not find a guarantee for the prevention of a substantive error, but we found a strongly mitigated probability of such an error occurring when no system error is found. Moreover there is a tendency that the size of the substantive error is mitigated by correct application of administrative procedures.
- On the other hand in 96% of the cases with a system error, no substantive error was found. This means that the systems assessment by the audit department is much more severe than necessary for the prediction of substantive errors. It is a consequence of the fact that the audit department includes many aspects in its systems assessment that are not directly relevant for a substantive error.

9.4 Predictability at the account level

For the predictability of the error at the account level, we had to aggregate the data of the individual transactions to the level of the account. At the account level the fraction of transactions containing a system error is a measure for the (lack of) quality of the operation of the administrative processes for this account. Therefore we use the rate of transactions showing a system error as the system error for that account. We refer to it as "fraction of system errors". We also compute the fraction of transactions in which a substantive error occurs: "the fraction of substantive errors", next to the "mean taint" for the account: the sum of all taints divided by the number of transactions in the sample from the account. With these fractions and means we can investigate predictability of the substantive error from the system error. We will do this for the relations as given in table 9.7. In this table the first column gives the predictor, the second (at the same row) the predictand, so sometimes the same predictor is used for different predictands (rows 1and 2).

Table 9.7: Relevant predictions

predictor	predictand	subsection
fraction of system errors (t)	fraction of substantive errors (t)	9.4.1
fraction of system errors (t)	mean taint (t)	9.4.1
fraction of subst. errors (t-1)	fraction of subst. errors (t)	9.4.2
mean taint (t-1)	mean taint (t)	9.4.2
fraction of system errors (t-1)	fraction of system errors (t)	9.4.2
fraction of system errors (t), plus mean taint (t-1)	mean taint (t)	9.4.3

The first 4 predictions of table 9.7 directly relate to the research questions 17.2 and 20 in 9.2.4; the letter "t" refers to a year and "t-1" to the year preceding the year "t". We added the last prediction, because that might be optimal: the information of the error size of last year combined with the information of the quality of the operation of the system of this year. It refers to research question 21.

9.4.1 Predictability substantive error from system error

When there is a high fraction of transactions containing one or more system errors, the account concerned may be expected to run a higher risk of showing substantive errors, in two ways:
by containing more transactions with a substantive error (the occurrence of substantive errors) or by showing a higher mean taint (the size of the error in the account).

Occurrence of substantive errors from occurrence of system errors
We start this subsection by looking at the relation between the occurrence of system errors and the occurrence of substantive errors, both in the year t, represented by their fraction. Table 9.8 gives the relevant Pearson-correlations.

Table 9.8: Fraction of system errors x fraction of substantive errors

year	P-correlation	R^2	n	p-value
1995	.207	.043	49	.15
1996	.366*	.134	43	.016
1997	.296*	.088	47	.044
1998	.316*	.10	50	.025
1999	.518**	.268	44	0

* significant at the .05 level **significant at the .01 level

Table 9.8 shows that the predictability of the occurrence of substantive errors, from the occurrence of system errors is satisfactory: two explained variances (column R^2) are larger than 10% (27% and 13%) and 4 out of 5 correlations are significant at the 5% level (at least). So the occurrence of system errors has predictive value for the occurrence of substantive errors.

Size of substantive errors from occurrence of system errors
The next question is whether the occurrence of system errors is also predictive for the size of the substantive error in an account. Therefore we compute the correlations between the fraction of system errors and mean taint per account. Because the size of the error is the decisive property in the audit opinion, we will both compute the P(earson)-correlation and the K(endall)-correlation, in order to get a robust analysis (see remarks in chapters 6, 7 and 8). The results are given in table 9.9.

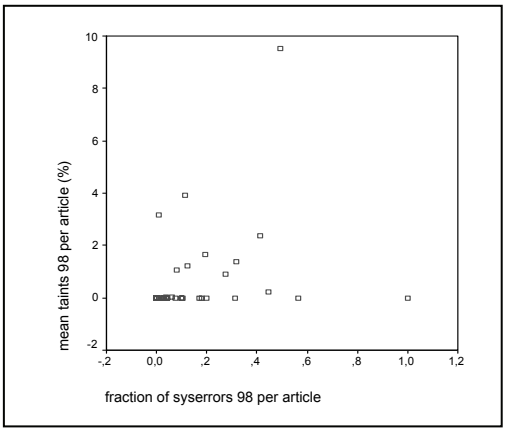
Table 9.9 Fraction of system errors x mean taint

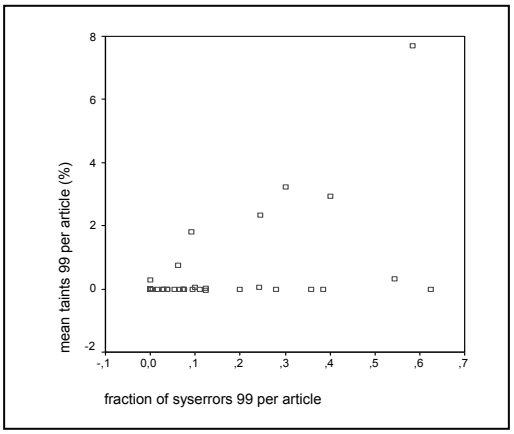
year	P-correlation	R^2	p-value	K-correlation	p-value	n
1995	.080	.006	.583	.242*	.024	49
1996	.167	.028	.285	.375**	.002	43
1997	.136	.018	.36	.148	.206	47
1998	.322*	.10	.023	.387**	.001	50
1999	.531**	.28	.0	.389**	0	44

* significant at the .05 level **significant at the .01 level

Table 9.9 shows two years (1998 and 1999), in which both the P-and the K-correlation are significant. These years also show the highest explained variances; only that for 1999 is greater than 10%. Scatterplots showed that the relatively substantial differences between the P-correlation and K-correlation for the years 1995 and1996 are not due to outliers. For the two strongest P-correlations found, we will show a scatterplot. in figures 9.1

Figures 9.1: Fraction of system errors against mean taint in 1998 and 1999





Both the scatterplots for 1998 and 1999 show one evident outlier. But they indicate that, on omitting this, a positive correlation will remain, which appears to be so. The P-correlations become:
for 1998: .153 (p-value .295, n=49) and
for 1999: .358, (p-value: .018, n=43).
So for 1999 the relation between system error and size of the substantive error is convincing and fully consistent with the high odds ratio found in 9.3.2. For 1998 the relation is less strong.

In summary, the most essential relation, that between the occurrence of system errors and size of substantive errors, is weaker than that between the occurrence of system errors and the occurrence of substantive errors. But for 1999 it is significant and for 1998 there is a fair indication of the relation to be expected with system testing having predictive power for the error rate.

Following table 9.2, we asked the question whether the quality of the operation of the system is a predictor for the trespassing of the 1% level for the mean taint. In order to come to an answer we again categorised the mean taints into the classes of table 9.2 and computed the mean fraction of system errors for each of these categories. Evidently the highest category of mean taints may be expected to coincide with the highest average of fraction of system errors; or rather when this coincidence occurs it is an indication that system tests predict the highest mean taints. We give the results in table 9.10.

Table 9.10: Mean fraction of system errors by taint category (M)

mean taint M (%)	1995 mean fraction	n	1996 mean fraction	n	1997 mean fraction	n	1998 mean fraction	n	1999 mean fraction	n
M<0	.19	5	.18	1	.14	3	.2	1	.12	1
M=0	.19	23	.12	26	.12	33	.76	37	.071	31
0<M<=.5	.24	15	.37	9	.20	4	.18	3	.23	6
.5<M<=1	.36	2	.10	1	.051	2	.28	1	.062	1
1<M	.32	4	.25	6	.26	5	.22	8	.33	5
Total		49		43		47		50		44

In 1997 and 1999, the highest category of taints coincides with the highest fraction of system errors. In 1995 and 1996 it coincides with the second highest and in 1998 with the third highest. Most striking is that by far the highest mean fraction of system errors is found for taints = 0 in the year 1998, for 37 cases!

The question was whether the fraction of system errors predicts the trespassing of the 1%-level of materiality. From this viewpoint we see that for all years the mean fraction of system errors is the highest (for 2 years), or relatively high, so to a certain extent the answer is positive. But a high fraction of system errors is not necessarily followed by a high taint.

Conclusion research question 17.2: quality of the operation system at the account level predictive for the fraction of substantive errors and/or mean taint?
In 4 out of 5 years the occurrence of system errors predicts the occurrence of substantive errors, in 2 out of 5 years the occurrence of system errors predicts also the size of the substantive error. So the most essential relation, that between the occurrence of system errors and size of substantive errors, is weaker than that between the occurrence of system errors and the occurrence of substantive errors.

Remark
The better prediction of the occurrence of substantive error than of the size, is consistent with our findings in 6.5.1 and 8.5.1 (tables 6.18 and 8.15).

9.4.2 Predictability error from the same type of error previous year

So far we have investigated the predictability of the substantive error by means of the system error. In this subsection we will investigate the predictability of the substantive error 'of this year' from the substantive error 'of the previous year' and next to that also the similar predictability of the system error. Finally we will investigate the predictability of the substantive error in year t from the substantive error in year t-1 and the system error in year t. The last predictability could perform best: it uses two sources of information which both are the best we can think of (except for the negative regression –effect). Analysis of our data will learn us whether we have to think of even better sources.

9.4.2.1 Predictability substantive error (t) from substantive error (t-1)
Because we had data of five years at our disposal, we could correlate the error with the error of the previous year, to look whether the latter could be used as predictor for the former. We investigated the predictability both for the fraction of substantive errors and for the mean taints. Table 9.11 gives the results for the fraction, i.e. the occurrence of substantive errors.

Table 9.11: P-correlations for the occurrence of substantive errors over years

years	fraction (t) x fraction (t-1)	R^2	n	p-value
1995-1996	.17	.029	39	.30
1996-1997	.52**	.27	41	.001
1997-1998	.31*	.094	44	.043
1998-1999	.49**	.24	37	.002

* significant at the .05 level **significant at the .01 level

We see that only the correlation for 1995-1996 is insignificant at the 5% level and we also see that the explained variance for two pairs of years is some 25%. This is a very satisfactory correlation.

Will this satisfactory level also apply to the correlations regarding the size of the errors in subsequent years? The answer is given in table 9.12.

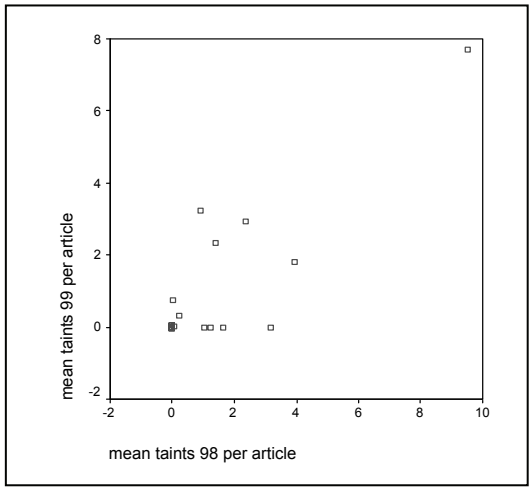
Table 9.12: P-correlations for the error size over years

years	Mean taint (t) x mean taint (t-1)	R^2	n	p-value
1995-1996	.084	.007	39	.61
1996-1997	.21	.045	41	.18
1997-1998	.047	.002	44	.76
1998-1999	.861**	.74	37	0

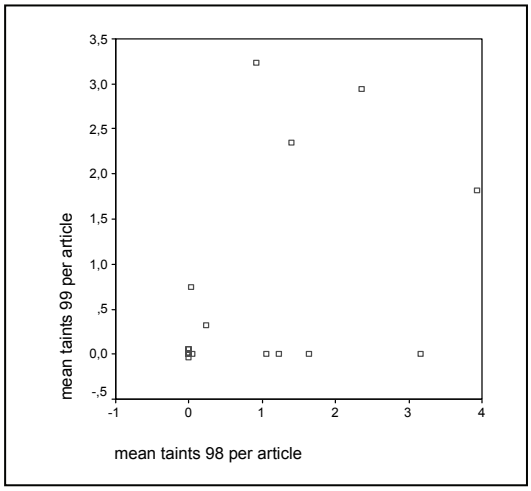
**significant at the .01 level

We investigated the high correlation for 98-99 by producing a scatterplot. It appeared that there is one obvious outlier; on omitting this, the second plot in figures 9.2 results. The skipping of the outlier clearly deflates the P-correlation, but it is still high(.54) and highly significant (p-value: .001). So the relation is convincing.

Figures 9.2: size 98 x size 99 *size 98 x size 99 (outlier omitted*



P-correlation:.86, p-value: 0 n=37



P-correlation: .54, p-value.001, n=36

Looking at all correlations in table 9.12, we see that for the size of the error the relation between the subsequent years virtually disappears, with an obvious exception for 98-99. So 3 of the 4 correlations are insignificant and small, but one is extremely large. An outcome that is hard to explain, albeit that the correlations in table 9.9 are also high for the year 1999. It looks as if (only) the errors of 1999 behave in a way we expect with regular system tests (table 9.9) and with a logical relation with the errors of the previous year (table 9.12). This regularity is also shown by the high odds ratio in table 9.5.

We could not find an explanation for this 'proper but deviating behaviour'[21].

Intermediate conclusion
We can conclude that predictive quality of the substantive error of the previous year for the same error of this year can be strong, both for occurrence and size, but like the validity we found in our first study, it is unpredictable if it really is. So it can not be relied upon.

[21] An explanation could have been that the negative regression (see section 8.7) does not occur because the maintenance for 1999 was more evenly spread. But we could not trace signs of such an even spread. And next to that, this would only be an explanation if a strong relation between system errors and mean taint would exist. But table 9.9 shows that this is not self evident, albeit reasonable for 1998 and 1999.

9.4.2.2 Predictability system error (t) from system error (t-1)
It is not the first objective of this research to investigate the predictability of system errors. But it makes sense to do that because, in risk analysis, both the error of the previous year and the quality of the administrative procedures of the previous are important. We compute the same P-correlations as we did in table 9.12, now for the fractions of system errors. The results are given in table 9.13

Table 9.13: P-correlations for the "occurrence" of system errors over years

years	fraction system errors (t) x fraction system errors (t-1)	n	p-value
1995-1996	.51**	39	.001
1996-1997	.23	41	.15
1997-1998	.22	44	.15
1998-1999	.855**	37	0

** significant at the .01 level

Two correlations turn out to be high and significant; all correlations are positive, which means that there is a certain predictability of the quality of the operation of the administrative procedures. The fact that only two correlations are significant indicates that the predictability is not very stable over years. All correlations are a counter-indication for the negative regression effect, because this would cause negative correlations.

To check the highest correlation, we show a scatterplot for the fractions of system errors for 1998 and 1999 in figure 9.3.

Figure 9.3: fractions of system errors 1998 x 1999

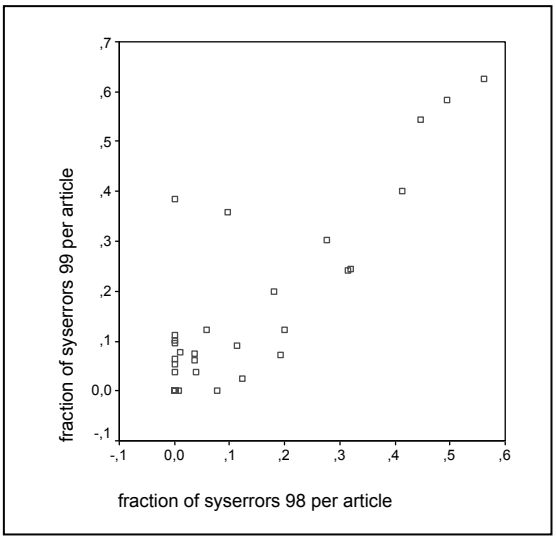


Figure 9.3 shows that the high correlation is not dependent on one or two outliers. So for this pair of years the correlation is very satisfactory.

From the viewpoint of control over the administrative processes, the picture is less satisfactory: high fractions of system errors should not tend to stay at a high level and high fractions should not occur where last year they were low.

Conclusion research question 20: errors (fraction, mean taint) year t-1 predictor errors (fraction, mean taint) year t?
The substantive errors of the previous year appeared to be predictive for the substantive errors of this year, but this predictability is not stable over the years; its stability is larger for the occurrence than for the size. The same lack of stability applies to the predictability of system errors from those in the previous year: in two years the relevant correlation is highly significant in the other two they are far from the 5% significance.

9.4.3 Predictability substantive error (t) from system error (t) and substantive error (t-1)

Probably the most logical dependence is the one in which the substantive error in year t is seen as predictable from the quality of the administrative system in year t and the substantive error in year t-1. We have investigated this relation by way of multiple regression. The dependent variable was the mean taint in an account, the independent variables were the fraction of system errors in year t and the mean taint in year t-1. We show the correlations between the two predictors in table 9.14 and give the results of the regression analysis in table 9.15.

Table 9.14: P-correlations fraction of system errors (t) x mean taint (t-1)

years	P-correlation	p-value
1995-1996	.21	.20
1996-1997	.12	.44
1997-1998	.16	.31
1998-1999	.40	.013

Table 9.15: mean taint (t) from fraction of system errors (t) and mean taint (t-1)

years	R	Adjusted R^2	F	p-value
1995-1996	.22	0	.995	.39
1996-1997	.31	.05	1.99	.15
1997-1998	.31	.05	2.14	.13
1998-1999	.89	.78	30.5	.000

With one exception, the relations found all are very weak, although the multiple correlation coefficient (R) is relatively high. But the adjusted R^2 is a better measure for the strength of the relation and this is very close to zero in three cases, corresponding to the high p-values found. Only the prediction for the error of 1999 is very satisfactory, which is consistent with the corresponding correlations in table 9.9 and 9.12. Table 9.14 shows that there is no collinearity to influence the regression results

Conclusion research question 21: errors (mean taint year t-1 and fraction system errors year t) predictor errors (mean taint) year t?
Only for the year 1999 the predictability of mean taint (t) is (more than) satisfactory. Prediction based on two predictors does not improve the prediction based on only the mean taint (t-1).

9.4.4 Conclusions on predictability on the account level

When we come to conclusions on predictability of substantive errors on the account level, obviously we have to make a distinction between their occurrence and their size.

The predictability of the occurrence of substantive errors turns out to be relatively good: the correlations found are of a satisfactory magnitude and significant at the 5% level or stronger in many cases. When we look at the size of the substantive error, we also find satisfactory correlations with the occurrence of system errors or with substantive errors of the previous year, but less consistently than for the prediction of the occurrence of substantive error. System tests are a not a stable predictor of the size of the substantive error.

This leads to the same warning as we gave for the use of risk analysis: only use system testing as a justification for the reduction of substantive testing when its validity in that respect has been established. Actually this warning is more serious than that regarding the use of risk analysis, because there we implicitly assumed that for organisations the

quality of risk assessment is stable over years. This 3rd study shows that for systems testing this stability can not be established. This raises new problems as to the certainty there is with regard to the possibility to rely on systems testing. Problems that might very well extend to the possibility to rely on risk assessment. We will discuss this further in chapter 10.

9.5 Discussion and conclusion

9.5.1 Discussion

We will discuss three possible causes for the sometimes absent, and sometimes relatively strong predictive power of system tests: (1) a logic that will cause this absence, (2) the way errors are distributed, (3) shortcomings in the definition of system errors.

9.5.1.1 An explaining logic?
When we look for explaining logic, we look for reasons why violations of administrative procedures do not necessarily affect the size of the error in an account. Especially when the procedures are meant to further the quality of the account, it will be hard to find such a logic, because logic was used to design the controls in such a way that they should mitigate the substantive error. An extra problem is, that we do not specifically know about the logic that led to the (violations of) controls, given in 9.2.2. Yet another problem is that the logic we have to find should not exclude satisfactory (1998) or even high (1999) correlations as we have also found. We found the general logic as can be found in Blokdijk (2004) to be a good starting point. And we will speculate on reasons why this logic still allows the higher correlations we found.

In his article Blokdijk discusses the effectiveness and efficiency of the various types of controls as part of risk analysis. To that end he distinguishes three stages of the preparation of financial statements. (1) the occurrence of events and their first recording in the accounting system (2) data processing, resulting in a routine product, "the trial balance" (3) adjusting the trial balance in order to arrive at the final balance sheet and income statement. In his opinion, in the third stage, only substantive procedures are fit for auditing. So the question of the effectiveness and efficiency of tests of control, system tests, only applies to the first and second stage.

In the first stage many so-called "non-reproducible controls" are relevant (controls that cannot be reperformed by the auditor, like the first recording of an event). His first conclusion is that non-reproducible controls should not be included in risk analysis, or in determining the control risk, because lack of compliance in the actual operation of non-reproducible internal controls cannot be remedied by the auditor's substantive procedures. But Blokdijk does not question the effectiveness of non-reproducible controls for the quality of the account, so here we do not find a starting point for the logic we are looking for. Moreover, the categorisation of system errors in 9.2.2 does not give information on which possible non-reproducible controls were considered in the dual purpose tests of this study.

For the second stage, that of data processing, Blokdijk discusses four types of general ICT controls: (a) controls of change management; (b) access controls; and application controls, being (c) programmed controls and (d) user controls. He also discusses the tests of control mentioned in ISA 400 par. 30. For all these controls he concludes that they have relevance for the quality of the accounts; he also concludes ‘that separate tests of control do not make much sense; if useful, their use lies in focusing substantive

tests rather than in assessing the internal control risk'. With that conclusion, together with his observation that in some situations the most effective tests of control are actually analytical procedures, he implies that tests of control do not contain much information on the substantive error. So, according to Blokdijk's logic, system tests can not be expected to have much predictive value for the error and maybe we should sooner be surprised by the presence of such predictive value than by its absence.

In evaluating the Blokdijk logic, we are heavily hindered by the fact that the specific definitions of the possible system errors, as given in 9.2.2, are not very clear about the precise nature of these possible errors. What, for instance, are "essential data", where are the "user controls", how influential for the error are the accounting rules, etc. It was practically impossible to find answers to these questions. We must conclude that this weakens the possibilities to conclude on the extent to which Blokdijk's logic applies to our data. But even if we are handicapped in that respect, we can conclude that our results do not falsify Blokdijk's logic, because relations between system errors and substantive errors are too frequently absent.

Implicit in the above discussion is that Blokdijk's logic does not exclude that better operation of the controls may coincide with less substantive error. It only implies that this coincidence might be caused by other factors than the quality of the operation of the system of controls. Plausible candidates for these other factors are psychological: attitude, motivation and the like. When an organisation's employees score highly at these properties, this will probably cause good performance on the operation of the controls and cause high quality accounting. When the employees score less highly, it still may be possible that the operation of the controls is satisfactory, but that this does not extend to the accounting. In this research we will not go deeper into these speculations, because we do not have the data to substantiate them.

9.5.1.2 The way errors are distributed

Figures 9.1 show error distributions on the Y-axis, implying very many zeros, many small errors and some larger errors. In fact in 1998 (1999) of 50 (44) mean taints 37 (31) were 0, 5 (5) smaller than 1% and the rest larger than 1%. The scatterplot of the errors in accounts for 1998 against 1999 also shows this (figures 9.2), The last figures also show that even with a relatively high correlation, the size of the error in 1999 can vary considerably for known error in 1998.

The error shows a pattern of noise at a very low level, with an occasional, but so far only weakly predictable outlier. Maybe we must conclude that the control measures in accounting work very well to keep the error almost always at an acceptably low level, but that in an individual case they do not predict how low exactly, albeit that their quality can be made high. Maybe it is like measuring a length: it is 99.7 cm, plus or minus 0.2 cm. Such a level of precision can only be outperformed by professionals. In auditing, the professionals are satisfied when the length is at least 99 cm, were 100 is aimed at, but where reaching that aim is very costly. Therefore we are satisfied with a production of the accounts of somewhat less precision. But then we should not be surprised if sometimes the outcome is even worse than the desired precision.

Accepting this rather fatalistic (realistic?) point of view of course has consequences both for the application of risk analysis and for the use of system tests. We go deeper into that in chapter 10.

9.5.1.3 Shortcomings in the definition of system errors in this study

We already mentioned the rather vague character of the descriptions of possible violations of compliance. They are not exclusive, it is very hard to weigh them for their seriousness. So the analyses we did, necessarily were of a crude nature. This may

have hidden a stronger relationship than we could find in some years, while in other years with the same labels, but maybe different content, the relations were stronger.

A remarkable property of the list in 9.2.2 is that it mostly regards aspects that have to do with legality and regularity, and only contains controls having to do with the accounting system itself at a rather high level of abstraction, like "compliance to accounting regulations". More concrete is "authorisation" and ‘errors in administration of obligations and maybe some aspects are hidden in "miscellaneous errors" that refer to the accounting system itself.

This observation leads to a further reflection on the generalisability of our findings. It is obvious that this generalisability is highly dependent on the way system properties and system tests are defined. If it were the case that in other audit departments the aspects of the accounting system itself prevail over other aspects, our findings probably would not generalise to these types of audit departments, which latter types will be found in the audit practice in private firms.

9.5.2 Conclusion

Three claims with respect to system tests are relevant for the audit practice:

1. They are supposed to give evidence on the truth and fairness of the financial accounts. Even if this would not be an explicit claim, it is implied by the fact that together with risk analysis that assesses low risk, less substantive testing may be done than without system tests (see ISA 400).
2. They are supposed to be fit for underpinning of a risk assessment "low", when this assessment serves as a justification for a decrease in substantive effort, compared to the effort needed in a situation without such a risk assessment.
3. They are supposed to give evidence in themselves on the quality, especially the operation, of the administrative procedures; they take the design as a given.

This part of our research was especially designed to give evidence on the first claim, operationalised as the predictive power for the substantive error of system tests. We give a closer look to the three claims mentioned.

As to the first claim, we saw the extent to which system tests give evidence on the truth and fairness from two viewpoints.

1 The predictive power for the occurrence of substantive errors at the level of a transaction was reasonable, although the relations found were not homogeneous over the years. There was also a predictive power for the occurrence of substantive errors at the level of the account.

2 The predictive power appears to be less, when at the level of the account the size of the substantive error is to be predicted from the quality of the operation of the system; for two years (98, 99) this predictive power could be shown. As the audit opinion is based on the size of the substantive error, it must be concluded that it is not self evident that system testing adds to the assurance regarding the audit opinion.

As to the second claim, the same lack of stability of predictive value for the size of the substantive error, leads to the conclusion that it is not self evident that system testing has the underpinning qualities it is claimed to have, as far as risk assessment is used as a justification for a decrease of substantive testing. This quality should be assessed, before an organisation relies upon the underpinning.

As to the third claim, the quality of system testing as direct evidence of the quality of the administrative procedures is only shown, insofar it proves to be a predictor of the

occurrence and size of substantive errors. The administrative procedures appear to have this quality, but unfortunately not stable over years, so it can not be relied upon.

Next to these three claims, it speaks for itself that a system test by nature gives evidence on the operation of the administrative procedures and the attitude and discipline of the employees involved. In that respect much more is relevant than just the question whether a substantive error occurs in a transaction. We mention for instance the possibility of omissions in the procedures regarding contracts to be acquired, and indications on how keen an organisation is in maintaining the quality of its processes. For an auditor these are very interesting aspects, basic for getting an overall picture of the organisation under audit.
This wider relevance is shown in our data by the fact that system errors occurred some 10 times as much as substantive errors. So naturally, the system test gives information on the quality of the administrative procedures and the extent to which they are maintained. The fact that they do not predict the size of the substantive error in an account leaves that unaffected.
An improvement might result from a division in the system errors into errors that have predictive power for a substantive error and errors that miss this power. For such an operation, the definitions would have to be sharp enough.

Chapter 10: Conclusions, Discussion and Perspectives

When auditors would consistently manage to give risk assessments close to the 'real risk', our validation results would have been better; nevertheless for some of the participating organisations the validation results are satisfactory. The trouble is that an organisation should be sure of the quality of its risk assessment before using it to warrant an unqualified opinion and our research shows that it cannot be, unless it has validated its own risk assessments. Apart from this conclusion, this chapter gives new possibilities to improve the practice of risk assessment, by giving it a firmer statistical basis, which even may result in coming very close to the 'real risk'.

10.1 Conclusions

10.1.1 Validity of the assessment of occurrence risk

In the first study, risk assessment showed a satisfactory validity with respect to two of the validation criteria: the error rate and the empirical distribution of the error rates. This validity was almost absent in the second study. The relatively favourable results of the first study have to be placed in perspective, because neither error rate nor empirical distribution incorporate the dependence of risk upon the level of materiality. Next to that, plausible moderator variables had no effect on the strength of the relation between assessed occurrence risk and error rate while the correlations found were not stable across organisations.

The other two validation criteria, 'audit position' and sampling risk, do incorporate the dependence of risk upon materiality. Validity with respect to these criteria was virtually absent, with an exception for two organisations.

Combining these varying indications for validity, we conclude that the assessment of occurrence risk can be valid, but that an audit organisation surely can not take this validity for granted. As the validity we investigated is necessary for the replacement of substantive testing by risk analysis, we formulate our first conclusion in the form of a warning.

The first conclusion:
Unless an audit organisation has sufficient evidence on the validity of its risk assessment, it should assume that this risk assessment is not valid in the sense that it may replace substantive audit.

10.1.2 Improvement by decomposition

We tried to accomplish improvement of the validity of risk assessment by its decomposition into the assessment on risk indicators. We used the error rate as a validation criterion, because it was the most promising where validity of the classical assessment of the occurrence risk was concerned. With this criterion we tried to find regression models in various ways, that would better predict the error rate. None of these approaches led to a systematic improvement of the prediction of the error rate, compared to prediction by the occurrence risk.

The second conclusion:
Improvement of risk assessment by decomposition of the occurrence risk with the help of the indicators in this study, could not be shown.

10.1.3 System tests as predictor of the error rate

We analysed the predictive power of system tests (tests of control) both for the occurrence and for the size of substantive error, as it showed itself in the data of more than 37000 dual purpose tests in the course of five years..

The third conclusion:
Fairly stable over years, the occurrence of a system error has a predictive power for the occurrence of a substantive error, both at the level of a transaction and at the level of an account.

The fourth conclusion:
The occurrence of system errors in an account has predictive power for the size of the substantive error in that account, but it varies considerably over years.

10.1.4 The "error of last year" as predictor of the "error of this year"

Both in our second and in our third study, we could analyse the predictive power of the size of the error 'of last year' for the size of the error 'of this year'. For the six instances we could analyse this predictive power was absent, except for one in the third study. In three of the four instances of the third study the occurrence of errors 'of last year' appeared to have predictive value for the occurrence errors 'of this year'.

The fifth conclusion:
The size of the substantive error 'of this year' is not predicted by the size of the substantive error 'of last year'. The occurrence of substantive errors of this year is relatively stably predicted by the occurrence of substantive errors in last year

10.2 Discussion

We concluded chapters 6 up to 10 with discussions on the results presented there. We will not repeat these, but will add some aspects resulting from the whole of the 3 studies. We will discuss (1) possible causes for the relatively unfavourable results, (2) compare the results to findings in literature, which are not consistently positive either, but still show stronger results, (3) the generalisability over the 3 studies, (4) the question whether we did accomplish what we intended.

10.2.1 Causes and solutions

We discuss four possible causes for the relatively unfavourable results.

(1) The first possible cause could be the *difficulties in general that man has with assessing risks*. In chapter 3 we discussed many possible biases from which this assessment may suffer. We also saw that many of the biases can be reduced by proper training and that in many cases practice furnishes this proper training. But we are not sure of the question whether this training was also received in the cases we studied. So here lies a possible cause.

The solution to this possible cause might be to provide for the desired training. But we wonder whether any training can be given in such a way that it really meets the practical conditions an auditor is engaged in. It would probably only work if designed for the specific way of assessment of an individual auditor: is he inclined to overstate risk, is he inclined to understate risk, does he need training in the branch of the business, or in the audit risk model, etc. However, as a consequence of Waller (1993) and Srinindi & Vasarhelyi (1986), there is one approach which might be successful for a great variety of contexts. Auditors could be trained to analyse more explicitly on the level of assertions and to aggregate the assessments per assertion to the level of the account in a proper way. We will further discuss this in section 10.3.

(2) A second possible cause could be the *level of difficulty which is inherent in predicting the error rate* in an account, due to its distributional properties. In almost all cases, the error rate is smaller than 1% (and then often 0, or much smaller than 1%). Most of the exceptions are only slightly larger than 1% and only occasionally really large. So we might be looking at administrative processes of which the outcome in terms of errors can be described as "noise". Predicting noise is very hard. This suspicion of noise and the difficulties in predicting it, is confirmed by the correlations of the error "of previous year" and the error "of this year", which are all close to 0 and insignificant (except one). And, as we could see from the scatterplots, an error of 0 in the previous year appears to be no guarantee for a zero or a small error this year; a large error in the previous year very often is followed by a small error this year. So, what might be expected to be one of the best predictors of the error rate, also turns out to be without predictive power. So we could wonder whether it is possible at all to predict the error rate from the qualities of the administrative processes and/ or the error of last year in an individual case.

The solution to this possible cause maybe can be found when we only reckon on the distributional properties of the error rate over a set of audit cases. When we know that error rates behave as noise around a low average level (which indeed they do in many cases); obviously we could use this as prior knowledge when starting a new audit. In section 10.3 we will discuss how we can take advantage of the distributional properties.

(3) A third possible cause might be that we had *"bad luck*" with the organisations that participated in the research. Or in other words, that these results are not representative of the way risk assessment is done in practice. We discussed this possibility earlier under the question of generalisability.

Obviously there is no solution to this possible cause, but we can put it into perspective:

- a perspective on the qualities of the selected organisations: we did not select them for their good or bad quality in risk assessment;
- a perspective on the reputation of the participating organisations: none was selected for either a good or a bad reputation with respect to the quality of their risk assessment. In fact this reputation did not play a role at all.

So we stick to our guess that the participating organisations perform neither better nor worse than the average audit organisation.

(4) A fourth possible cause might be the *regression effect* in the assessment of risk we introduced in 3.2.1 expectation E. This regression could be on the findings of last year: when these were (1) favourable, the assessment of the occurrence risk is more likely to be favourable too; when these findings were (2) alarming, an assessment ‘high’ for the occurrence risk is more likely, even if the administrative processes were improved.

When the first regression effect is combined with the unpredictability of the error rate, as we found it in chapters 8 and 9, this explains ineffective audit designs (‘under-

auditing'): the extent of substantive testing or other substantive audit is chosen too low, given the actual risk. Because the actual error and therefore the actual risk, given a low error rate in last year, can be 'high' just as well as 'low' this year. This regression in risk assessment would be unintended and would lead to 'under-auditing'.

When the second regression effect is combined with the unpredictability of the error rate, as we found it in chapters 8 and 9, this explains inefficient audit designs ('over-auditing'): the extent of substantive testing or other substantive audit is chosen too high, given the actual risk. The second regression effect also combines with a natural course of action in auditing and accounting. In this course, the auditor urges the ones responsible for the accounting to improve their procedures and/ or operation. Most probably the accounting department follows the relevant advice and thus improves the administrative processes. By way of system testing and other means of systems assessment, the auditor sees that the processes are of higher quality than last year. So the auditor is fully justified to assess the occurrence risk at low(er), but in many audit practices it is still customary to choose for OR='high'', as a consequence of a 'safety first' principle. In this case, the regression is intended and the auditor actually did the right assessment. He just did not act according to this assessment, for he also gave much weight to his assessments of the previous year.

From the viewpoint of the quality of an audit this course of action is fully justified, but it disturbs our view on the capabilities of an auditor with regard to risk assessment; actually he is better at it than can be read from his audit files. Many auditors with whom we discussed our findings, confirm this practice of 'safety first'. In future research this regression effect is worth to be investigated. We will include it in our discussion on the perspectives in section 10.3.

So of the four possible causes the last one may be applicable, the third one is not expected to be effective, the second one cannot be remedied, but circumvented as we will show in section10.3 and the first one may be true and a proper training may be available (see also section 10.3).

10.2.2 Comparison with results from literature

Much of the research discussed in chapter 3 does not directly regard the relation between risk assessment and error rate. In principle only the literature that refers to consistency of risk assessment with the error rate, is comparable to our study. We discussed

1. Roberts & Wedemeyer (1988); they found consistent relations between six general attributes, comparable to control environment, and the size to be expected of the error.
2. Asare & Davidson (1995); they found three studies in which risk factors predict the error rate and two in which they do not.
3. Waller (1993) investigated risk assessment at the assertion level; he found a positive correlation between the assessment of inherent risk and the rate of qualified opinions in a set of 215 real-life audits.
4. Kreutzfeldt & Wallace (1990), found positive correlations between the error rate and many from 75 operational variables describing the control structure
5. Wallace & Kreutzfeldt (1993), found positive correlations between the error rate and five factors regarding the control structure.
6. Wright (1994) found increasing incidence and impact of errors with decreasing strength of controls.
7. Bell & Carcello (2000), found that logistic regression on six risk factors improves prediction of fraud

The first, fourth, fifth and sixth study do not regard the risk assessment as such, but relevant aspects for the assessment. In the studies these aspects, called "factors" or "operational variables" have a rating on their quality and with this only implicitly on the risk associated with them (high quality => low risk). Under this condition all studies show consistently positive relations between factors and error rate, where in the fourth study these positive relations regarded ca. 80% or more of the investigated variables. This outcome is not fully comparable to the risk assessment studied in this thesis, because the link to an assessed risk is not made and therefore a crucial aspect of risk assessment is missing.

The findings in our first study on the relation between risk assessment and error rate are also relatively favourable, although they varied considerably over organisations. When we include the findings of our second study, the results deteriorate and the overall picture becomes that of an unreliable relation between risk assessment and error rate.

The studies referred to, and our studies, have in common that they investigate real-life situations, so are fully comparable on that aspect. Maybe the most important difference between the studies is that in our study the occurrence risk is quantified, whereas it is not in the referred ones; there the assessment only regards aspects of the occurrence risk. And this may account for the difference in the results.

This supposition is confirmed by some of the findings of the second study, that of Asare and Davidson, where 5 studies are reviewed. In 2 of them no relation is found between risk assessment and error rate. But in 3 studies they do find this relation. We actually did two studies into the validity of the assessment of the occurrence risk. In the first the validity on the error rate was satisfactory for a couple of organisations, but for the remaining organisations and for those participating in the second study, validity with respect to error rate was absent. With these findings we can conclude that our results are similar to those of Asare & Davidson, but with more weight on the absence of validity.

Waller's study investigates only the relation between inherent risk and error rate, which makes the results not comparable. Still it does show that risk assessment at the assertion level does relatively well.

So our studies may show less strong relations between risk assessment and error rate but these differences can be explained by the difference in handling risk assessment. This conclusion leads to a perspective: maybe risk assessment should only be done at the assertion level. This perspective is further discussed in section 10.3.

The sixth study, that of Wright, is comparable to our study into system tests. Wright also went into the strength of the controls, where we only looked at the operation of them given their strength. Maybe that difference accounts for the difference in outcome between his and our study. Both studies find a relation between controls and incidence of errors, but Wright also finds a relation with impact of the error, where we find a varying relation between the occurrence of system errors and the size of substantive error. Maybe these differences can also be accounted for by our observation that the definitions of the system errors in our study could have been sharper, thus excluding the possibility of rating them for their impact.

In the study of Bell & Carcello the effectiveness of decomposition is investigated. They found predictability of fraud and improvement from decomposition. But this result is only partially comparable to ours, because their assessment regards the occurrence of fraud, not its size and moreover, fraud has a fundamentally different relation with the

administrative procedures than most of the other misstatements that can occur in an account.

10.2.3 Generalisability over the three studies

We mentioned it earlier: the availability of our data may have influenced what we found. This influence in our first study, may have been caused by the willingness of the organisations to participate in our research. It may also have played a role in the choice of the cases that were reported via the questionnaires, but we may expect that for the respondents the retrieval of the data from their audit files will have been relatively easy In our second study we ended with less departments than we started, because with some departments the necessary data were only available at too large an effort. This may have influenced what we found.

In both studies it is more likely that it will have influenced the relation in a positive rather than in a negative direction. For available transparent data are likely to coincide with audit of a high quality. On the other hand, the availability was in the relation to our questions; with other questions the availability might have been fully different. So the possible influence can only be speculated at, but we expect it to be negligible.

The availability was evident in our third study: we just were given the opportunity to analyse a host of interesting data. It hardly makes sense to speculate on the extent to which the audit department we got the data from, is representative of other audit departments. A sample of size one (organisation) can never be strong as a basis for generalisation. Still more than in the other studies of this thesis, the generalisation will be that of "existence" (or “to a theory”, (Yin, 2003), see also section 5.6). The strength of the data in the third study lies in their big numbers.

For the three studies as a whole these considerations mean that a rigorous generalisation is not possible. The results should be seen as strong indications for the (lack of) validity of risk assessment. They give a strong warning not to assume that risk assessment is valid by itself. They must be seen as a less strong indication that tests of control, or system tests might not work as well as they are expected to work.

We can also put it another way: if both risk assessment and system testing would properly work, we most probably would have found stronger relations than we actually did.

10.2.4 Did we accomplish what we intended?

In chapter 3 we concluded that the heuristics and biases paradigm is interesting, but research based on it does not give conclusive information on the validity of risk assessment in practice. Similar conclusions could be drawn from the consistency studies. Our review showed very interesting things on the conditions under which consistency in the assessments will be seen and/or will indicate more quality of risk assessment. However no validation was performed on the error rate or on an estimated audit risk, except for a couple of studies.

So we concluded that there was a gap to be filled and that in this thesis the filling of this gap would be started.

With the conclusions of 10.1, we may claim that we indeed started this work, or, with a view on the six studies we mentioned in the start of this section, continued it. But we do

not claim that we have filled the gap: there are still many questions unanswered. Some of them will be treated in 10.3.

10.3 Perspectives

We discuss 4 perspectives for further research we believe to be promising for improvement of risk assessment, given the results of this thesis:

(1) risk assessment on assertions rather than at the account level
(2) predicting the error size instead of assessing the occurrence risk
(3) in combination with (2): another way of eliciting and representing risk
(4) taking the empirical distribution of the error rates as a point of departure.

We will discuss them in the same order

10.3.1 Risk assessment on assertions rather than at the account level.

For the convenience of the reader we refresh our discussion of the term 'assertion'. In 3.3.3 we explained that the truth and fairness of the annual accounts is made up of aspects like existence, accuracy and completeness. These aspects are labelled 'assertions' (also 'audit objectives', or 'controle criteria' in Dutch); so assertions regard the aspects of truth and fairness, as they are subject of an audit. The term 'assertion' derives from the form in which they appear: "The account of debtors is complete" has the form of an assertion. Truth and fairness of the account implicitly or explicitly consists of such assertions. Actually you can not talk of truth and fairness without defining the aspects, assertions, that make up this property. Not all assertions are applicable to all sub-accounts.

Waller (1993), Srinindi & Vasarhelyi (1986) and Buckless (1989) indicate that risk assessment at the assertion level may lead to better results than risk assessment at the account level, provided a suitable aggregation rule is available. This possibility is consistent with the general idea of improving assessment by reducing complexity by decomposition and the idea in particular of our study that risk assessment might be improved by decomposition of the assessment task into assessment on assertions (see 3.3.3). This consistency is not taken away by the lack of improvement we had to face with our indicator approach, because the indicators imply a decomposition into risk factors, whereas the assertions imply a decomposition into aspects of truth and fairness (see also figure 3.1 in 3.3.3).

When we suppose that risk assessment on the assertion level leads to better assessments per assertion, the problem of a proper aggregation rule is still left. Srinindi (1984) gives such aggregation rules, Vrijling & Van Gelder (2002) also give a way to deal with the aggregation of separate risks in the context of hydraulic engineering and controlling the various risks concerning budget execution in projects.

This possible improvement for risk assessment implies research in which two randomly selected groups of auditors do their risk assessment, one group by means of classical risk assessment at the account level (so only with implicit assessment on the assertions), the other by means of explicit assessment of risks at the assertion level; assessments on the assertions are aggregated to the level of the account in the two ways referred to above. The outcomes of these risk assessments are validated on the error rate, on the sampling risk and on the conditional distribution of the error rates, provided enough cases can be generated.

10.3.2 Aim at prediction of the error rate and not at assessing risk

In this thesis we have identified two ways of prediction of the error rate to be expected as outcome of the audit:

(1) in our first study we got a strong indication that risk assessment in the classical way as such is better in predicting the (size of the) error, than in assessing the risk (of missing a material error); in the first and second study there were indications that risk assessment is fairly good at predicting the existence of errors.
(2) various studies we referred to in our review of the literature succeeded in locating factors, more or less representing aspects of the control structure, that are good predictors of the error found in the corresponding account.

We think that especially the second approach, in which the quality of aspects of the control structure is assessed, without explicitly attributing an associated risk to them, could lead to improvement of risk assessment, because the judgmental part of this procedure has less weight than in risk assessment. So we may circumvent the problems that assessment of risks in the audit field meets. Some of our references, summarised in 10.2.2, indicate that we may expect improvement.

Both in accordance with the first approach (risk assessment predicting error), and in accordance with the second approach (factors control structure predicting errors) 'risk assessment' could be aimed at assessing which size of the error is to be expected, given the quality of the administrative processes (and inherent risk factors). An audit will produce an estimate of the error. This study would be aimed at the predictive qualities of the assessments according to the first or to the second approach. A convenient property of a study that would aim at investigating this quality, may be that it can combine both approaches.

Assessing a most likely error (for the account) is not the same as assessing a risk, and it is not self-evident how this can be integrated in the calculations of the reliability of the audit. We have to extend this procedure in order to come to an outcome of this "risk assessment" that really has the properties of a risk. This extension is given in 10.3.3.

10.3.3 Another way of modelling prior knowledge.

Once the auditor has assessed the most likely error that will be in the account, he may wonder how sure he is that the actual error will not be "far" from this expected error. He can do this by stating (assessing) probabilities on the expected and on error rates nearby and more distant. These probabilities will be dependent on the outcome of the procedure (linear regression, logistic regression, etc.) in which he predicted the error. This procedure will have some statistical measure that indicates a bandwidth around the expected error. And this bandwidth relates to the probability that can be given to this expected error. On the other hand he also may use his assessments from the classical risk analysis and guess how probable that expected error will be, compared to other possible error rates. He may also combine the statistical and judgmental approach.

This attribution of probability to the expected error will also be dependent on the level of refinement with which the other possible errors are viewed. To complete his attribution of probabilities, the auditor will define a suitable grid of possible error rates with values 0 up to 1 (included). In most situations he will give the error rates less than 1 percent (or another small percentage, associated with the level of materiality) almost all probability to be distributed, and the error rates larger than 1% only a small part of it. But of course, this will depend on the error(s) he expects as a consequence of his risk analysis or of the procedure meant in the second approach in 10.3.2.

Acting in this way, the auditor attributes the probabilities by way of what mathematicians call a “discrete distribution”: probabilities are assigned to a discrete set of (possible) error rates p (and not to a continuously changing set of p’s, which would lead to infinitely many possible error rates).

The auditor could attribute the probabilities in many other ways, for instance by assigning to each possible error rate (of the grid he has chosen) an equal probability (1/N if a grid of N points is chosen). This gives a uniform distribution on [0,1], the interval for the error rates p and thus expresses that he has no prior knowledge on the error rate (or does not want to make use of it). Or by giving a greater weight to small error rates (and less to the larger ones), thus expressing great confidence in the quality of the account, or by giving a greater weight to larger error rates, thus expressing more reservations about the quality of the accounts; virtually every prior picture of the possible p’s and their probability can be modelled with this simple tool. This way of modelling prior knowledge has many advantages:

(1) it may combine quantified methods to estimate the most likely error to be found in the audit with more judgmental part of risk assessment;
(2) it meets the findings in our study where the validation criterion at which the risk assessments showed most quality, was the error rate, which, as mentioned before, points at an assessment which reflects ideas about the error rate;
(3) it offers more flexibility in modelling than continuous models of known distribution families. As a matter of fact in auditing these models often appear to be too strict, thus causing for instance the necessity of assigning a separate discrete probability to an error rate of zero (or one, or both); in fact the proposed way offers the possibility to model any mode, also multiple modes;
(4) it is much more informative than the present practice: not only the prior risk of a material error is given, or directly computable, also the extent to which the auditor is certain of his/her assessments is reflected: the more certain that the error rate will be some p_0 or within narrow boundaries from p_0, the higher the probabilities assigned to p_0 and the other p-values within the boundaries;
(5) it offers quite natural possibilities for validation: the error rate in the sample validates the mode of the prior; also the likelihood of the error rate, based on the sample results directly validates the prior (except for one serious complication: the dependence of the sample size and provided the prior probabilities do not deviate too much from the beta distribution, if that is the distribution with which the sampling risk is calculated);
(6) it circumvents the necessity of justifying a mathematical model for the distribution of the error rate;
(7) the occurrence risk in this method is very easy to be assessed: it is the sum of the probabilities attributed to the error rates equal to materiality or larger;
(8) when the auditor chooses for the ‘All or Nothing’ approach for the evaluation of the audit sample, implying that a monetary unit either is in error (for 100%), or is correct (for 100%) and he uses monetary unit sampling, the binomial law applies to the distribution of the number of errors in the sample; this offers simple ways to combine the occurrence risk and the sampling risk by means of what is known as the Extended Bayes Rule[22]

[22] This rule reads as follows: Let $\{H_i\}_{i=1,...,n}$ be a set of hypotheses about a parameter p; let $P(H_i)_{i=1,...,n}$ be a prior distribution on H_i; let $P(x|H_i)$ be the probability of x, given H_i; then $P(H_i|x)=P(x|H_i)P(H_i)/\Sigma\ P(x|H_i)P(H_i)$

(9) this way of modelling the prior information can also fully or partly be based on the empirical distribution as is discussed in the next subsection.
(10) this way of modelling the prior information and jointly analysing sample results with the Extended Bayes Rule circumvents the statistical invalidity of the Audit Risk Model.

Next to these convenient properties, the approach has the limitation that no simple statistical evaluation method is available when the more common taints approach is chosen for the evaluation of the sample results.

Performing risk assessment in a combination of the ideas of 10.3.2 and 10.3.3 can very well be supported by research: the method of assessing the error to be expected can directly be validated. The method of eliciting the occurrence risk, as given in 10.3.3, can be validated with the sampling risk as a criterion or with the distribution of the error rates conditional on the assessed occurrence risk in a number of audit cases.

10.3.4 Taking the empirical distribution of error rates as a point of departure.

In chapter 2 we gave a justification for the use of the empirical distribution of the error rates of a set of audit cases, conditional on the level of occurrence risk, as a validation criterion for the occurrence risk. The ideas developed there can easily be extended into a way of dealing with prior information, both from statistical analysis of known error rates in a set of "suitable" cases and from the assessment of occurrence risk in the audit being performed.

We take S to be a set of "suitable" audit cases. In a number of these cases an error rate has been observed, meant to lead to an audit opinion on p, the error rate in the account of the case under audit. We have the empirical distribution of these error rates, say E. If we may assume that a new audit case A is a member of S, randomly drawn from it, we can see E as the prior distribution of the error rate in A.

For the validity of this approach, it is crucial how we define ‘suitable’. A tautological definition would be “‘Suitable’ is what makes the set fit as the basis for the prior distribution”. But this ‘definition’ still gives an idea how to construct a suitable set.

The most practical approach in constructing a suitable set is just to start with all cases an audit department/ organisation audited last year. In principle, their empirical distribution is fit for being used as prior distribution of the error rate in the account of a new audit case, provided it may be seen as a randomly selected member of the set. This set may be refined by taking subsets on some of the criteria just mentioned. The stability requirement will prevent audit cases from being ‘suitable’ when they are too old.

If the empirical distribution E appears to vary with the level of the assessed occurrence risk, say that distributions E_{vl}, E_l, E_m, E_h are found for the assessed occurrence risks, respectively ‘very low’ (VL), ‘low’ (L), ‘medium’ (M), and ‘high’ (H), and we have assessed an occurrence risk for A, we can refine the prior distribution for A into say E_l, if the assessed occurrence risk was ‘low’ and likewise for a possible other occurrence risk.

If we do this, a practical programme for continuing research on the validity and possible improvement of risk assessment unfolds, in which from the start, advantage with respect to the efficiency of the audit can be taken from information on the error rates in the near past and in which risk assessment can be validated:

(1) A start up situation can be created, by establishing the "suitable" set from the near past of completed audit cases. These cases all showed some estimated error rate. The empirical distribution of these error rates is modelled. This can be done either by adopting the strategy of 10.3.3 (where probabilities were assigned to a discrete set of possible error rates), or by adopting some known family of distributions, for instance the beta distribution (as we did in 6.5.1 and 8.5.1).
(2) Dependent on the number of suitable cases that can be found and on the assessed risks in these suitable cases, it is worthwhile to model the empirical distribution separately per level of assessed occurrence risk for the same suitable cases.
(3) With a couple of adaptations the empirical distribution (either per level of OR, or irrespective of OR) of the "suitable cases" can be used as the prior distribution for the audit of this year, and may serve as a basis for the calculation of the number of necessary substantive tests. We will discuss these adaptations also as a separate research topic.
(4) The audit is performed, assessment of the occurrence risk included, an estimated error rate is established. Evidently materiality, occurrence risk and the estimated error and other key features of the audit are kept record of.
(5) For all audits that can be seen as a member of the suitable set of audit cases, the procedure of the previous point is executed; this results in a data set consisting of pairs "occurrence risk, estimated error rate".
(6) If wanted, more variables can be kept record of and measurements added to the data set of the previous point; variables like size of the account, attributes of the auditor, attributes of the audit object, etc.. Evidently the choice of these variables depends on the planned research topics.
(7) The generic research question: "Do the distributions of the error rates differ for varying assessed levels of occurrence risk?" now can be answered by analysing these distributions for the various levels of occurrence risk as assessed in the set of audit cases of this year. This approach allows a non-monotonous relation between OR and error rate, which makes the outcomes more realistic and informative.
(8) Other questions should be answered too:
is the "error of last year" predictive for the "error of this year"?
is the distribution of the "errors of this year" stable compared to the distribution of the "errors of last year"? Both questions are relevant because of the "suitability" of the starting set.

The idea of this approach is that the information contained in the prior distribution of the error rates in the "suitable set", is in itself relevant for the reliability of the audit opinion aimed at. Therefore it may serve as a basis for reduction in the extent of substantive testing, even without making use of risk assessment.

It may even contain more information than risk assessment. This is indicated by the results of the first study in this thesis. Here we found the distributions that fit to the distribution of the error rates for the total of the 120 cases and for the cases for varying occurrence risk. For this we calculate

(1) the size for a sample of substantive tests, if we would use this "suitable set" regardless of the information given by the risk assessment, so by using the set of 120 cases and
(2) the size for a sample of substantive tests while taking into account that the empirical distributions differ after the level of assessed risk.

We calculate the size of the sample by designing an audit for a materiality level of 1% in which the substantive tests lead to an unqualified opinion as long as no more than 1

error is found in the sample (we choose a "k=1 sample"). If we would not use prior information, a sample size of 475 would be needed.

When we use the property of the beta-binomial distribution as given in property 4.1 in chapter 4 and use the results of 6.5.1, we get the sample sizes as given in the next table.

Table 10.1: Sample sizes in needed for various prior distributions

suitable set	# of cases*	(a,b) prior beta distribution	Sample size needed	less audit effort
all 120 cases	120	(.164, 9.77)	321	154
cases with OR=VL	9	(.186, 25.34)	309	166
cases with OR=L	35	(.127, 38.71)	285	190
cases with OR=M	58	(.388, 38.18)	332	143
cases with OR=H	17	(.729, 9.40)	419	56

* 120 error rates were recorded; in one case the assessment of OR was missing

Indeed table 10.1 shows that the largest gain in terms of less audit effort is got from the fact that prior information from the empirical distribution of the "suitable set" is used; a reduction in the audit effort of 154 individual transactions is justified. For three of the levels of the occurrence risk the audit effort differs less than 37 individual transactions; only when OR=H, 98 extra transactions have to be audited. But note that in the classical approach the auditor would audit 475 individual transactions in this case.

Note that table 10.1 is just used as a hypothetical example. In practice adaptations as discussed below in 10.3.4.2 should first be made.

The outcomes of table 10.1 make clear how improvement of risk assessment may affect the necessary audit effort: the better the risks are assessed, the larger the differences will turn out to be for the sample sizes still needed. The approach of this subsection turns out to offer the organisation that will engage in it the opportunity to "earn" a greater differentiation in the sample sizes needed, corresponding to the improvement in risk assessment. The research part of this approach simply is to assess the conditional distributions for the various levels of assessed risk. It can be extended into various other directions, discussed now.

10.3.4.1 "Suitable sets"

A crucial research topic in this approach will regard the question: "What makes sets suitable?". The generic answer to this question is relatively simple: A set is suitable if its members have sufficient in common with the new audit cases for which a prior distribution is needed.

This may mean that all audit cases of one audit department of the previous year can be considered to be ‘suitable’. But if the audit cases of this year contain large deviations from those of last year, there will be a problem. It is hard to decide when this deviation will be too large. Decisive criterion for this question is that the distribution of possible error rates for the new audit case may be expected to be the same as that of the error rates in the suitable cases. This criterion gives a way to investigate the suitability of the set of cases that furnishes the prior distribution.

The question would be which properties of an audit case make that it complies to the decisive criterion.

- The properties could regard type of transactions: expenses regarding personnel, activities regarding the creation of infrastructure like the construction of roads,

bridges, railways, etc., they may apply to the transfer of money in the form of subsidies, loans, etc..
- The properties could also regard the stability of the audit object, its size, its context, etc.. When major changes in the business and/ or accounting processes have taken place since an error rate was observed, such an audit case will be less fit for inclusion in the suitable set.
- The properties could regard the type of business: retail, wholesale, hotel and catering, etc., stability, size and many other properties may be important.

It is obvious that relevant properties for audit objects from the public sector will differ from those in the private sector.

10.3.4.2 Adaptations of the prior distribution based on the suitable set

A second research topic, related to the previous one, regards the question which adaptations of the empirical distribution, based on the suitable set, are necessary to let it be fit as the prior distribution for a new audit. This topic arises from the observation that a new audit case only can be seen as a random selection from the suitable set, if this represents the potential audit cases for the new year. So audit cases of this year and audit cases of last year have to be interchangeable. In order to assess this interchangeability it is necessary to assess factors determining this interchangeability. Many of them will be a form of stability, on the properties mentioned in the previous paragraph. The extent to which this stability can be made plausible determines the extent to which the suitable set can be thought representative for the potential audit cases of this year. Depending on the extent of representativeness, still adaptations of the empirical distribution of the error rates of the suitable set, or of conclusions based on this, will be necessary.

We can think of the following adaptations:
inflate the necessary size of substantive testing as results from the plain suitable set, by some factor, dependent on a minimum for substantive testing that is deemed to be necessary, the similarity of the audit cases with the suitable set, the time lag between the audit cases and the suitable set, etc. These adaptations can be tested for their necessity when a database is built over years, in which adaptations and their causes are recorded.

10.3.4.3 Regression effects

When in the classical approach of risk assessment the auditor willingly introduces regression effects as discussed in 10.2.1(4), he might consider not to do this when making use of this empirical distribution as a prior, because the empirical basis enhances the reliability of this prior information, compared to judgmental risk assessment. If the auditor keeps opting for a ‘safety first’ approach, it is wise to keep record in which cases OR=’high’ was chosen in spite of the awareness of a quality of the administrative processes that would allow to assess a lower level for OR. Such a record allows to treat deliberate risk avoiding risk assessment separately when risk assessment is validated.

Of course the approach of this subsection gives no clues as to how to improve risk assessment. But it can be combined with the possibilities given in the first three approaches. It may be tempting to observe that this improvement is not really necessary for the goal of decrease in the audit effort. This goal can already be attained by making use of the statistical properties of the distribution of the error rates in the suitable set.

Our choice from the four possibilities mentioned in section 10.3 for the continuation of research and improvement of risk assessment, would be the last one: the one based on the empirical distributions, because

1. to a considerable extent, it circumvents the difficulties we find in predictability of the error rate, either by risk assessment or by the error of the previous year;
2. it starts with the probable bonus that, regardless of the quality of risk assessment, decrease of the audit effort can be justified and that on a sound statistical basis;
3. this choice does not keep us from introducing elements from the other three approaches.

10.3.5 Best Practice Research

Improvement of risk assessment could also be found by way of best practice research. Organisations' risk assessments are validated, organisations showing assessments with highest validity are compared for possible determinants for their best practice. We will not further discuss this possibility.

10.4 Business process analysis

Since about a decade, the audit risk approach is gradually being replaced by the business risk approach, or business process analysis/ approach (BPA). We discussed this in 1.2.7. As we mentioned there, our research exclusively deals with the audit risk approach, simply because our data apply to this approach. But this exclusive orientation leaves us with two related questions:

1. Do the results of this study also apply to risk analysis in BPA?
2. Would the results with respect to the validity of risk assessment have been better, when BPA had been used, instead of the ARM?

As to the first question, we firstly have to add "as far as it also aims at assessing the occurrence risk". We mentioned that the Joint Working Group (JWG 2000, p.10) has the same expectation as to the effectiveness and efficiency of BPA as have Van Leeuwen and Wallage (see Van Leeuwen & Wallage 2002), especially for issues concerning accounting estimates, going concern and management fraud. But the Joint Working Group also notes that the auditor will have to transform outcomes of the business process analysis as to their relevance for the quality of the accounting processes and for the inherent risk. There might be a possibility that ICR and IR are assessed with more quality, because of the wider scope of BPA. But it is not very likely that this possible improvement will be significant, because that would imply that the methodology of the ARM would insufficiently consider the relevance of the business processes for the audit risk. But the opposite is true: the ARM explicitly aims at incorporating all aspects of the business relevant for the quality of the administrative processes, both the uncontrollable (in the inherent risk) and the controllable in the control risk). So the improvement that may be expected from BPA with respect to the assessment of OR results from its wider scope. Our conclusion with respect to the first question is, that we expect our findings to apply to BPA, as far as it is aimed at assessing OR.

As to the second question, we observe that our analysis of and answer to the first question implies our answer to the second question: we do not expect a significant improvement from the application of BPA for the assessment of the occurrence risk.

Obviously these two expectations ask for substantiation by empirical research.

Glossary

In this glossary definitions of key concepts and some of their semantics are given. Concepts in a definition that refer to a defined concept in this glossary are in italics.

1. *Allocation:*
   primary allocation: the choice on which balances or classes of transactions the auditor will deploy his audit resources.
   secondary allocation: the determination of the extent of the audit activities per chosen *audit object*
2. *Analytical procedures*: analyses meant to detect signals in or produce signals from the data that underlie or constitute the account under audit; signals that point at the possibility of errors in the account.
3. *Assessment:* finding out what could be the value of a relevant attribute, by making use of a mix of more or less precise methods, which may vary from judgmental methods, intuition included, analysis of processes of key features, comparison with similar situations, transformation of qualitative findings into quantities, to strictly quantitative measurements. The mix chosen is also a matter of judgment.
4. *Annual accounts*: a set of accounts, issued every year, reporting the state of the finances of a business. The annual accounts often are made up by sub-accounts. In the case of a business the sub-accounts mostly are debtors (receivables), creditors (payables), stock, assets, loss and profit.
   In this thesis the audited account often will be a sub-account. In case of a governmental organisation, the sub-accounts may be obligations, payments, receipts, balance.
   sub-account: a group of homogeneous individual items, of which the aggregate is included as a separate item in the annual accounts, or in management information that is summarised in the annual accounts
5. *Assertion:* 'truth and fairness' of the annual accounts are made up of aspects like existence (of debtor, stock), accuracy (correctness of a booking), legality (of a transaction). These aspects are called 'assertions'. Audit establishes truth and fairness with respect to these assertions, also called 'audit objectives' (or, in Dutch: 'controle criteria')
6. *Audit assurance (AA):* the assurance or level of reliability that is associated with an unqualified opinion; it is the complement of the audit risk (AA=100%-AR)
7. *Audit object*: the entity of which the audited annual accounts have to give a true and fair view. The audit opinion essentially regards the accounts, but the audit implies examination of all the relevant persons, procedures, authorizations, administrative organization and internal controls, next to contextual factors like type of business and quality of the personnel. "Audit object" can refer to a business or organisation as a whole, but also to parts of the business or organisation, with its own processes and sub-accounts.
   This means that the audit object is much broader than the audited accounts.
8. *Audit objective*: See *assertion*.
9. *Audit risk (*AR): the risk that the auditor gives an *unqualified opinion*, when in fact the audited accounts contain a *material error*.
10. *Audit risk approach*: the approach to auditing in which the analysis of risks as to the existence and detection of a material error plays a central role in the design and execution of an audit and in which the *audit risk* is modelled by way of the *audit risk model*.
11. *Audit risk model*: the model in which the *audit risk* is seen as a function of *the inherent risk*, the *control risk,* the *risk of analytical review* and the *sampling risk*.

12. *Audit value*: the value which, according to the audit, a transaction should have. This value results from the audit of the *book-value* and underlying transactions.
13. *Bayesian statistics*: a statistical methodology in which information on the parameter of some probability distribution, for instance of errors in an account, is modelled as a probability distribution on this parameter and inferences are on the probability distribution of this parameter.
    In this methodology, a natural sequence of events is:
    a) an investigator has information on this parameter, for instance: "The mean error rate per monetary unit in this account may be expected to be very low";
    b) this information is modelled in a "prior distribution" on this parameter;
    c) the investigator draws a sample from the distribution, for instance from the errors in the account (by selecting a sample of *line items* and assessing the error per *line item*) and assesses the probability of the sample result;
    d) the investigator combines the prior probability and the probability of the sample result into a "posterior distribution" for the parameter, the mean error rate in our example.
    Conclusions about the parameter (again: the mean error rate, in our example) are based on the posterior distribution, in which, for instance, the probability that the mean error exceeds materiality can be calculated. (see Lee,1997 pp 33,34)
    In this methodology the prior information may be based on previous statistical investigations, but it is also allowed that the prior information is an expectation, a state of belief or a conviction. In this way it is distinct from what is called "classical statistics". In the same way it is similar to what is done in the *audit risk approach*.
14. *Book-value:* the monetary value at which a transaction is recorded; in the account, such a record is also called 'line item'.
15. *Business process analysis:* the approach in which the whole of the context of the business and all of its processes, the administrative processes included, are analysed as to the risk that the business will fail to reach its goals.
16. *Compliance error:* failure to comply to the accounting and internal control system.
17. *Compliance testing:* see *tests of control.*
18. *Confidence level:* the probability that the estimation procedure results in an interval that contains the parameter that is estimated.
19. *Confidence upper limit*, also *confidence upper bound:* see *interval estimation*
20. *Conjunctive event*: an event that happens if (and only if) all of a set of composing sub-events happen simultaneously (are in conjunction); it is analogous to a series connection.
21. *Control risk (CR):* Control risk is the risk that a misstatement, that could occur in an account balance or class of transactions and that could be material individually or when aggregated with misstatements in other balances or classes, will not be prevented or detected and corrected on a timely basis by the accounting and internal control systems (ISA 400 par.5).
22. *Detection risk (DR):* Detection risk is the risk that substantive procedures will not detect a misstatement that exists in an account balance or class of transactions and that could be material, individually or when aggregated with misstatements in other balances or classes. (ISA 400 par 6).
    The detection risk can be broken down into:
    risk of analytical review (RAR): The risk of analytical review is the risk that analytical procedures will not result in the detection of a misstatement that exists in an account balance or class of transactions that could be material, individually or when aggregated with misstatements in other balances or classes (*ISA 520 par 10-15),*
    sampling risk (SR): Sampling risk arises from the possibility that the auditor's conclusion, based on a sample, may be different from the

conclusion reached if the entire population were subjected to the same audit procedure." (*ISA 530 par. 7*).

23. *Disjunctive event*: an event that happens if at least one of a set of composing sub-events happens; it is analogous to a parallel connection.
24. *Dual purpose test*: a test on an individual transaction in which both the operation of the controls (the administrative system) is checked and the correctness of the *book-value* is investigated.
25. *Error*: at the level of a *line item*: the difference between *book-value* and *audit value* (see also *misstatement*.); at the level of an account: the difference between the total of *book-values* and the total of *audit values*
26. *Error rate*: the total error amount divided by the total of book-values (= the size of the account).
27. *Inherent risk* (IR): Inherent risk is the susceptibility of an account balance or class of transactions to misstatement that could be material, individually or when aggregated with misstatements in other balances or classes, assuming that there are no related internal controls. (ISA 400, par. 4)
28. *Internal control risk:* see control risk
29. *Internal controls*: the set of measures aimed at proper functioning of the administrative processes.
30. *Interval estimation*: an estimation procedure in which the estimate takes the form of an interval, a range of possible values for the entity that is estimated. In auditing the Stringer bound is an example of an interval estimation: it gives the upper bound for possible error rates in the account. Because this upper bound is associated with a level of confidence, it is called 'confidence upper bound'; in auditing also the term '*Upper Error Limit' (UEL)* is in use.
31. *Likelihood:* There are two ways of interpreting a value of the formula for a probability distribution with some parameter p, for some outcome X:
    1. p is seen as the variable, and X as given; in this interpretation, the value of the formula is the <u>likelihood</u> of p (given X)
    2. X is seen as the variable and p as given; in this interpretation the value of the formula is the <u>probability density</u> of X (given p),
32. *Line item:* see *book-value*.
33. *Material error*: an *error* or *misstatement* in the account of a size that may influence decisions that have the *annual accounts* as a basis.
34. *Misstatement*: the difference between *book-value* and *audit value* (see also *error*).
35. *Monetary unit (MU)*: the unit or currency in which the book-value is recorded; also (used as) unit of sampling (See also *Monetary Unit Sampling*)
36. *Monetary Unit Sampling (MUS):* A sampling method in which the distinct monetary units (MU's) are the unit of sampling, each drawn with equal probability. So with N monetary units, each has probability 1/N to be drawn This causes a book-value to be selected with a probability proportional to its size.
37. *Most likely error (MLE):* see *point estimation.*
38. *MU-sample*: sample of monetary units, a sample in which every single monetary unit is drawn with equal probability (see also *Monetary Unit Sampling)*.
39. *Nominal level* is the level of reliability (or assurance, or risk) that was aimed at by the audit design; it is the level of reliability at which the size of substantive testing is calculated (by using a model, such as the ARM). In a valid model, this nominal level will be equal to the actual reliability.
40. *Occurrence risk*: the risk that the account an auditor has to audit, contains a material error. The occurrence risk can be broken down into *inherent risk* and *internal control risk*. Because of the problems to separate the IR and CR in an unequivocal way, many auditors do not separately assess IR and CR, but assess the occurrence risk OR as the combination of both. This means that OR = f(IR, CR).

41. *Point estimation:* an estimation procedure in which one single value is attributed to the entity that is estimated. This estimation in principle is 'best' with respect to a statistical criterion, for instance that the estimate has maximum likelihood, given the data in the sample. In auditing the sample mean of the taints is often used as an estimator. This meets the maximum likelihood criterion and therefore is also called *'most likely error (MLE)'.*
42. *PPS*-sampling: sampling of the units with a probability proportional to their size; *MU-sampling* effectuates PPS sampling of (the book-values of) transactions.
43. *Risk of analytical review* (RAR): see *detection risk.*
44. *Sampling risk* (SR): see *detection risk.*
45. *Simple event*: an *event* that is not structured as a combination of sub-events.
46. *Sub-account:* see *annual accounts.*
47. *Substantive tests of details*: substantive tests of details are part of substantive procedures: procedures concerned with amounts aimed at obtaining audit evidence to detect material misstatements in financial statements. Substantive tests verify one or more assertions about a financial statement (for example, the existence of accounts receivable), or make an independent estimate of some amount (for example, the value of obsolete inventories on the individual bookings in the audited account). (ISA 530 par 17).
48. *Taint (tainting)*: The amount of MU's in error divided by the book-value in a transaction.
49. *Tests of control:* based on the auditor's understanding of the accounting and internal control systems, the auditor identifies the characteristics or attributes that indicate performance of a control, as well as possible deviation conditions which indicate departures from adequate performance. The tests concern the absence or presence of these attributes (ISA 530 par. 15)
50. *Unqualified opinion*: An *unqualified opinion* should be expressed when the auditor concludes that the financial statements give a true and fair view (or are presented fairly, in all material respects) in accordance with the identified financial reporting framework . (ISA 700 par. 27)
51. *Upper error limit (UEL)*: see *interval estimation*.

Literature

Abdolmohammadi, M.& A. Wright(1987): *An Examination of the Effects of Experience and Task Complexity of Audit Judgments,* The Accounting Review, Vol LXII, No 1 January 1987.

Amer, Tarek Karl Hackenbrack & Mark Nelson (1994): *Between-Auditor Differences in the Interpretation of Probability Phrases* Auditing, A Journal of Practice and Theory (Spring), Vol. 13 no 1 (pp 126-136)

Arens, A.A. & J.K. Loebbecke (1997): *Auditing, an integrated approach (7th Edition)*. Prentice Hall, London.

Asare, S.K. & Davidson, R.A. (1995). *Expectation of Errors in Unaudited Book Values: The Effect of Control Procedures and Financial Condition.* Auditing, a Journal of Practice and Theory, Vol 14 No. 1.

Ashton, R.H. (1974): *An Experimental Study of Internal Control Judgments.* Journal of Accounting Research, 12 pp.143-157.

Ashton, R.H. & J. Kennedy (2002): *Eliminating Recency with Self-Review: The Case of Auditors' 'Going Concern' Judgments.* Journal of Behavioral Decision Making 15 pp. 221-231.

Babbie, E. (1994): *The Practice of Social Research (7th edition).* Wadsworth Publishing Company, London.

Bar-Hillel 1979. *The role of sample size in sample evaluation.* Organizational Behavior and Human Performance, 9, pp. 245-257.

Basu, Progyan & Wright, Arnold (1997): *An Exploratory Study of Control Environment Risk Factors: Client Contingency Considerations and Audit Testing Strategy,* International Journal of Auditing Vol 1 No 2 (pp 77 - 96)

Batenburg, Paul C. van, A. o'Hagan & R.H. Veenstra (1994): *Bayesian discovery sampling in financial auditing: a hierarchical prior model for substantive test sample sizes*. The Statistician Vol. 43 No.1 pp 99-110.

Bedard, J. (1989): *Expertise in Auditing: Myth or Reality?* Accounting Organisations and Society, Vol 14 No's 1/2 (pp 113-131).

Bell, T.B. & J.V. Carcello(2000): *A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting.* Auditing: A Journal of Practice and Theory Vol. 19 No. 1 (pp 169-184)

Bell, T.B., Knechel, W.R., Payne, J.L & J.J. Willingham (1998): *An Empirical Investigation of the Relationship between the Computerisation of Accounting Systems and the Incidence and Size of Audit Differences.* Auditing: A Journal of Practice and Theory Vol. 17 No. 1 (pp 13-38).

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975): *Discrete Multivariate Analyses: Theory and Practice.* MIT Press, Cambridge, Massachusetts.

Blokdijk, J.H. (2001): *De effectiviteit van de systeemgerichte aanpak in de accountantscontrole.* Maandblad voor Accountancy en Bedrijfseconomie Jg. 75 No. 3, maart (in Dutch).

Blokdijk, J.H. (2002): *De toetsing van de werking van de interne controle.* Maandblad voor Accountancy en Bedrijfseconomie Jg. 76 No. 11, december (in Dutch, pp.613-616)

Blokdijk, J.H. (2004): *Tests of Control in the Audit Risk Model: Effective? Efficient?* International Journal of Auditing. Volume 8 No 2 Page 185-194.

Bonner, S. E. (1999): *Judgment and Decision-Making Research in Accounting*. Accounting Horizons Volume 13 No 4 pages 385-398.

Box,G.E.B. & G.C. Tiao (1992): Bayesian Inference in Statistical Analysis. Wiley and Sons, New York.

Broeze, G.B., W.M. Lammerts van Bueren en F. Hartsuijker (1991): Audit risk Model: Toepassen bij Gebrek aan Beter? De Accountant Jg 98, nr1 (pp. 27- 32, in Dutch)

Broeze, G.B., N.G. de Jager & M.C.A. van Zuijlen (1997) *Risicobeoordeling op basis van indicatoren.* MAB (72e jaargang no 10, pp 539-546, in Dutch)

Buckless, Frank A. O. N.(1989): *Modelling External Auditors Evaluations of Auditrisk and the Effects of the Task Environment on Consensus*, Dissertation.

Burgstahler, D. & J. Jiambalvo (1986): *Sample Error Characteristics and Projection of Error to Audit Populations.* The Accounting Review Vol. 61, No. 2

Burgstahler, D., S.M Glover & J. Jiambalvo (2000): Error Projection and Uncertainty in the Evaluation of Aggregate Error *Auditing: A Journal of Practice and Theory Vol*. 19 No. 1

Butler, S.E. 1986. Anchoring in the Judgemental Evaluation of Audit Samples. *The Accounting Review* Vol. LXI no. 1. .

Cohen, J. & P. Cohen (1983): *Applied Multiple Regression/Correlation Analysis for the behavioural Sciences (2nd Edition).* Lawrence Erlbaum Associates, Publishers, London.

Colbert, Janet L. (1988)*: Inherent Risk: an Investigation of Auditors Judgments*. Accounting Organisations and Society, Vol 13 (pp 111-121).

Colbert, Janet L. (1989)*: The Effect of Experience on Auditors Judgments*. Journal of Accounting Literature, Vol 8 (pp 137-149).

Cook, T.D. & D.T. Campbell (1979): *Quasi- Experimentation, Design and Analysis Issues for Field Settings*. Rand McNally College Publishing Company , Chicago

Cushing, B.E. & S.S. Ahlawat (1996): *Mitigation of Recency Bias in At Judgment: the effect of Documentation*. Auditing, A Journal of Practice and Theory Vol. 15 No. 2 (pp 110-122).

Davis, E.B., S.J. Kennedy & L. Maines (2000): *The Relation between Consensus and Accuracy in Low-to-Moderate Accuracy Tasks: An Auditing Example.* Auditing, A Journal of Practice and Theory Vol. 19 No. 1 (pp 101-121).

Dusenbury, R., J.L. Reimers &S. Wheeler (1996): *An Empirical Study of Belief-Based and Probability-Based Specifications of Audit Risk.* Auditing, A Journal of Practice and Theory Vol. 15 No. 2 (pp 12-28).

Elder, R.J. & R.D. Allen (2003): *A Longitudinal Field Investigation of Auditor Risk Assessments and Sample Size Decisions.* The Accounting Review Vol. 78 No. 4, pp. 983-1002.

Ellifsen, A., W.R. Knechel & P. Wallage (2001): *Application of the Business Risk Audit Model: A Field Study,* in: Accounting Horizons, Vol. 15 no 3, September, pp 193-209.

Emby, C. 1994. *Framing and presentation Mode Effects in Professional Judgement: Auditors' Internal Control Judgements and Substantive Testing Decisions.* Auditing, a Journal of Practice and Theory, Vol. 13, Supplement .

Floor, K. (2002): *Tien jaar ensemblevoorspellingen op het Europees weercentrum.* Zenit, december (in Dutch)

French, Simon (1988): *Decision Theory, an Introduction to the Mathematics of Rationality,* Chichester Horwood.

Gaumnitz, B.R., T.R. Nunamaker, J.J. Surdick & M.F. Thomas (1982): *Auditor Consensus in Internal Control Evaluation and Audit Program Planning.* Journal of Accounting Research Vol. 20 No. 2 (pp 745-755).

Groeneboom, P. (1993) ``*Afvoertoppen bij Lobith",* work done by order of the Ministry of Public Works).

Hair, J.F., R.E. Anderson, R.L. Tatham & W.C. Black (1998): *Multivariate data Analysis*. Prentice Hall, New Jersey

Hall, Thomas W.,Terri L. Herron, Bethane Jo Pierce & Terry J. Witt (2001): *The Effectiveness of Increasing Sample Size to Mitigate the Influence of Population Characteristics in Haphazard Sampling.* Auditing: a Journal of Practice and Theory Vol. 20 No. 1

Handbook IAASB 2004: *Handbook of International Auditing, Assurance and Ethics Pronouncements* IFAC 2004

HCDAD (1997): *Handboek Controle DAD* (in Dutch)

Hendriks, H., C. Kraaikamp, L. Meester, P. Mokveld & M. Nuyens (2005): *Isolating and Correcting Errors while Auditing Accounts.* in: *Proceedings of the 48th European Study Group Mathematics with Industry* (C. Kraaikamp, H.X. Lin & C.W Oosterlee, eds.) Delft University Press, Delft.

Hogg, R.V. & A.T. Craig (1970): *Introduction to Mathematical Statistics*, Macmillan, New York.

Hoogewoning, P. (1991): *Het Audit Risk Model (ARM)*. De Accountant Nr. 1 pp. 36 – 38 (in Dutch)

Houston, Richard W. Michael F. Peters & Jamie H. Pratt (1999). *The Audit Risk Model, Business Risk and Audit-planning Decisions.* The Accounting Review,Vol. 74 no. 3 pp 281-298.

IAASB: *Handbook of International Auditing, Assurance, and Ethics Pronouncements*.

ISA 200: *Objective and General Principles Governing an Audit of Financial Statements* IFAC 2004.

ISA 330: *The auditor's procedures in response to assessed risks.* IFAC 2004.

ISA 400: *Risk Assessments and Internal Control.* IFAC 2004 .

ISA 520: *Analytical Procedures.* IFAC 2004

Johnson, R. (1987): *Auditor detected errors and related client traits - A sample of U.K audits.* The Journal of Business, Finance and Accounting 14 (Spring): pp 39-64

Johnson, R. 1988. *Inherent and control risk evaluations and the identification of account balances containing material errors*. Proceedings of the SixthUSC Audit Judgment Symposium. University of Southern California, CA.

Johnson, P.E., Jamal, K. & Berryman, R.G. 1994. *Effects of Framing on Auditor Decisions*. Organizational Behavior and Human Decision Processes 50 pp. 75-105.

Johnstone D.J. (1995): *Statistically Incoherent Hypothesis Tests in Auditing.* Auditing, A Journal of Practice and Theory Vol. 14 No. 2(pp. 156-175).

Joyce., E. (1976): *Expert Judgment in Audit Program Planning.* Studies on Human Information Processing in Accounting; supplement to Journal of Accounting Research Vol 14, pp 29-60.

Joyce, E.J. & Biddle, G.C. 1981a. Anchoring and Adjustment in Probabilistic Inference in Auditing. *Journal of Accounting Research Vol*. 19 No 1. pp 120-145

Joyce, E.J. & Biddle, G.C. 1981b. Are Auditors Judgements Sufficiently Regressive?. *Journal of Accounting Research Vol.* 19 No. 2 pp. 323-349.

JWG 2000. Recommendations arising from a study of recent developments in audit methodologies of the largest accounting firms. *Joint Working Group of* International Auditing Practices Committee, the Assurance Standard Board of the CICA, the Auditing Practices Board of the United Kingdom and Ireland and the Auditing Standards Board of the AICPA.

Kahneman, D. (2003): *A Perspective on Judgment and Choice, Mapping Bounded Rationality*. American Psychologist Vol 58 No. 9 pp 697-720.

Keeney, R.L. & H. Raiffa (1976): *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley & sons, New York

Kinney, W. R. (1992): *The audit risk model at the financial statement level:the joint occurrence risk problem*. Working paper, University of Texas.

Klijnsmit, P. M. Sodekamp en P. Wallage (2003):*Bedrijfsrisico's van de accountant en het Audit Risk Model.* MAB mei, 77e jaargang nr 5 pp 190- 195.

Kok, C.J. (2001): *Calibration of EPS derived probabilities.* Scientific Report; WR 2001-04 KNMI, De Bilt.

Koning, F. de (2002):*Beoordeling van de interne controle in het kader van de accountantscontrole.* MAB juni, 76e jaargang nr 6 pp 272- 280 (In Dutch).

Kreutzfeldt, R. & W. Wallace (1986). *Error chracteristics in audit populations: Their profile and relationship to environmental factors.* Auditing: A Journal of Practice & Theory 6 (Fall): 20-43

Kreutzfeldt, R. & W. Wallace (1990): *Control Risk Assessments: Do They Relate to Errors?* Auditing: A Journal of Practice & Theory 9 (Suppl.): pp. 1-26

Krug Nelson, M. 1995. Strategies of Auditors: Evaluation of Sample Results. *Auditing: A Journal of Practice & Theory* 14 (Spring): 34-49.

Kuijck, J.R.H.J. van (1999): *Control of Judgment Performance in Auditing*. Bob van Kuijck, Amsterdam (dissertation).

Lammerts van Bueren (1991): *Misschatting bij Accountants* De Accountant pp.33 -35 (in Dutch).

Lea, R.B., S.J. Adams & R.F. Boykin (1992): *Modelling of the Audit Risk Assessment Process at the Assertion Level within an Account Balance.* Auditing, A Journal of Practice and Theory Vol. 11 Suppl (pp.152-179).

Leadbetter, M., G. Lindgren. & H. Rootzen (1983): *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.

Lee, P.M. (1997): *Bayesian Statistics, an Introduction*, Arnold, London.

Leerboek Accountantscontrole (2003), Stenfert Kroese, Houten (in Dutch)

Leeuwen, O. van, Ph. Wallage (2002): *Moderne controlebenaderingen steunen op interne beheersing.* MAB maart 76e jaargang nr 3, pp 82-89.(in Dutch)

Loebbecke, J.L. (1995): *On the use of Bayesian Statistics in the Audit Process.* Auditing, A Journal of Practice and Theory Vol. 14 No. 2(pp.188-192).

Meixner, W. F. & R.B. Welker 1988. Judgment consensus and auditor experience. An examination of organizational relations*. The Accounting Review* 63 (July) 505-513.

Messier, William F., Steven J. Kachelmeier & Kevan L. Jensen (2001): *An Experimental Assessment of Recent Professional Developments in Nonstatistical Sampling Guidance.* Auditing, a Journal of Practice and Theory Vol. 20 No. 1

Mock, T. & A. Wright 1993. An exploratory study of auditor evidential planning judgments. *Auditing: A Journal of Practice & Theory* 12 (Fall): 39-61.

Mock, T. J. & A. Wright (1995): *Audit Evidential Planning: Further Archival Evidence on whether Programme Plans are Risk-Adjusted.*

Mollema, K. (2003): *Auditrisico, meer dan ooit een issue!* MAB, december, 77e jaargang nr 11(in Dutch).

Mollema, K. (2004): *Auditrisico, meer dan ooit een issue!* MAB, januari/ februari, 78e jaargang nr 1 (in Dutch). NIVRA geschrift nr 49 (1989): Accountantscontrolerisico (In Dutch).

Novick, M.R. & P.H. Jackson (1974): *Statistical Methods for Educational and Psychological Research*. McGraw-Hill, New York.

POB (2000) The Panel of Audit Effectiveness of the Public Oversight Board: *Report and Recommendations*, August 2000.

Reimers, Jane, Stephen Wheeler &Richchard Dusenbury (1993): *The Effect of Response Mode on Auditors Control Risk Assessments.* Auditing: a Journal of Practice and Theory (Fall) Vol. 12 No. 2.

Roberts, D.M. & P.D. Wedemeyer (1988): *Assessing the Likelihood of Financial Statement Error in Using a Discriminant Model,* Journal of Accounting Literature Vol 7 nr. (pp 133-146)

SAS No.39: *Audit Sampling.* AICPA 2001

SAS No.47: *Audit Risk and Materiality in Conducting an Audit.* AICPA 2002

SAS No.55: *Consideration of the Internal Control Structure in a Financial Statement Audit.* AICPA 1998

SAS No 82 currently SAS No.99: *Consideration of Fraud in a Financial Statement Audit.* AICPA 2004.

Schilder, A. (1991): *Risico-analyse: Rekenmodel, Denkmodel of Wedstrijdmodel? Een nabeschouwing bij een boeiend debat*. De Accountant Jg 97, nr. 9 (pp. 573 – 576, in Dutch).

Sennetti, J.T. (1995): *On the Incoherent Use of Evidence: Why Subjective Bayesian Evidence is not held Probative.* Auditing, A Journal of Practice and Theory Vol. 14 No. 2(pp.193-199).

Siegel, S. & N.J. Castellan (1988): *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, New York.

Smith & Kida (1991): *Heuristics & Biases: Expertise and Task Realism*. Psychological Bulletin Vol 109 No. 3, pp 472-489.

Snel, Bianca (2002): *The Stringer bound using the beta distribution as the posterior distribution (*graduation thesis*),* University of Utrecht.

Srinindi, B.N. (1984): *Probability Modelling of Internal Control Systems in Audit in Decision-making* Ph.D. Dissertation, Columbia University Graduate School of Business

Srinindi, B.N. & M.A. Vasarhelyi (1986): *Auditor Judgment concerning the Establishment of Substantive Tests Based on Internal Control Reliability.* Auditing, A Journal of Practice and Theory Vol. 5 No. 2(pp.64-76).

Stevens, J. (1996): *Applied Multivariate Statistics for the Social Sciences (3rd ed.)* Lawrence Erlbaum Associates, Publishers, New Jersey.

Stone, Dan N. & William N. Dilla (1994): *When numbers are better than Words: The joint Effects of Response Representation and Experience on Inherent Risk Judgments.* Auditing, A Journal of Practice and Theory Vol. 13 Supplement (pp.1-19).

Tamura et al.(1989) *Statistical Models and Analysis in Auditing.* Statistical Science Vol. 4, pp 2-33.year).

Tan, H-T (1995): *Effects of Expectations, Prior Involvement, and Review Awareness on Memory for Audit Evidence and Judgment.* Journal of Acccounting Research Vol 33 No 1 pp 113-135.

Tan, H.P., T.B. Ng & B. W. Mak (2002): *The Effects of Task Complexity on Auditors' Performance: The Impact of Accountability and Knowledge.* Auditing: A Journal of Theory and Practice Vol. 21 No. 2

Thorndike, E.L. (1920): *A Constant Error in Psychological Ratings.* Journal of Applied Psychology, Vol.4 pp. 469-477

Touw, P. & L. Hoogduin (2002): *Statistiek voor Accountancy.* Academic Service, Schoonhoven (in Dutch).

Trotman, K.T. & Wood, R. (1991): *A Meta-Analysis of Studies on Internal Control Judgments*. Journal of Accounting Research, Vol. 29 No.1.

Tversky, A. & D. Kahneman (1971): *Belief in the Law of Small Numbers.* Psychological Bulletin, Vol 76, No 2 pp. 105-110.

Tversky, A. & Kahneman, D. (1974): *Judgement under Uncertainty: Heuristics and Biases.* Science Vol. 185 No. 4157, pp. 1124 - 1131.

Tversky, A. & D. Kahneman (1981): *The Framing of Decisions and the Psychology of Choice.* Science, Vol 211 No 4481, pp. 453-458.

Van, G.H., R.J. Grübel & G.B. Broeze (1996): *De Studentized bootstrap methode in de accountscontrole* Maandblad voor Accountancy en Bedrijfseconomie, jg 70 nr 4. (in Dutch)

Veenstra, R.H. & P.C. van Batenburg (1990): *Enkele nadere beschouwingen over risico-analyse.* MAB, (pp. 102-104, in Dutch)

Vrijling, J.K. & P.H.A.J.M. van Gelder (2002): *Probabilistic Design in Hydraulic Engineering* TU Delft, Subfaculty of Civil Engineering, Delft.

Wagenaar, W.A.(1977): *De Beste Stuurlui dempen de Put.* Ambo (in Dutch).

Wallace W.A. & Kreutzfeldt, R.W. (1993): *The Relation of Inherent and Control Risks to Audit Adjustments*. Journal of Accounting & Finance pp. 459-481.

Waller, W.S.(1993): *Auditors' Assessments of Inherent and Control Risk in Field Setting*s. The Accounting Review Vol. 68 No. 4 pp 783 803.

Waller, W.S., & Zimbelmans, M.F. (1996): *Evaluating and Improving the Accuracy of Auditors' Risk Assessments.* Universities of Arizona and of Oklahoma, internal note.

Wijbenga, J.H.A., J.J.P. Lambeek, E. Mosselman, R.L.J. Nieuwkamer & R. Passchier (1993): *Toetsing uitgangspunten rivierdijkversterkingen.* (in Dutch) Ministerie Verkeer en Waterstaat.

Wilks, T.J. (2002): *Predecisional Distortion of Evidence as a Consequence of Real-Time Audit Review.* The Accounting Review, Vol. 77 No 1, pp. 51-71.

Willemsen, A. (1996): *On determining an upper confidence bound for the total error amount in audit populations.* Technical University of Delft, Limperg Instituut, Netherlands Court of Audit.

Willingham, J. & W. Wright 1985. *Financial statement errors and internal control judgments*. Auditing: A Journal of Practice & Theory 5 (Fall): 57-70.

Wolde, J. ten (1989): *Enkele Beschouwingen over Risico-analyse*. MAB, (pp. 331-340, in Dutch).

Wright, A. (1994): *The Relationship between Assessments of Internal Control Strength and Error Incidents, Impact, and Cause.* Boston College.

Yin, R. (2003): *Case study research; design and methods* (3rd edition). Sage publications, Thousand Oaks.

Appendix 1: The Questionnaire

The next pages show a prototype of the questionnaire we used in our first study. It is almost uniform for the various participating organisations, but some questions were adapted to the actual practice of this organisation. This adaptation regarded, for instance, the categories for the questions 1.1 and 1.7; for some organisations some risk indicators were added, others deleted. As explained in our thesis, the actual results were on indicators that were applicable to all organisations.

Contact persons, currency and other elements that could increase the recognisability of a participating organisation have been removed or masked.

The questionnaire had more questions than used in this thesis. All questions have played a role in the research report that was written for every participating organisation giving all results for this organisation. In our thesis we left out some of the aspects covered by the questionnaire.

Obviously for the Dutch organisations and firms the questionnaire was in Dutch.

The following explains how we came to risk indicators.

Desirable qualities of risk indicators.
Use of risk indicators in the assessment task will possibly improve risk assessment when

- A risk indicator represents a clear, recognizable aspect of the audit object.
- A risk indicator represents an optimal part of the total of relevant aspects: not too small, because then the interaction between indicators becomes too complicated, not too large because then complexity is not reduced enough.
- The set of risk indicators covers all relevant aspects.
- A risk indicator may be expected to be predictive for the error rate.
- A risk indicator is as independent as possible of the other indicators defined. “Independence” in this context is to mean: one can vary without the others necessarily covarying, for instance: educational level can vary independently of separation of duties, which only to a minimal extent can (or should) vary independently of authorisations.

These qualities were thought to be desirable, partly because of logical reasoning, partly because of the analysis of audit files in a pilot study (Broeze et al, 1997) and partly because of standard audit theory (see among others Arens & Loebbecke, 1997). They gave guidance when we constructed the set of indicators used in this thesis. The qualities were used as giving some direction in the construction of the set of indicators to be used in our research. This research will give some answer to the necessity or sufficiency or desirability of these properties for the indicators that were developed.

Construction of the set of indicators
In our study we used 20 risk indicators. An initial set was constructed in a pilot study (see Broeze et al, 1997). In this study five real audit cases on stocks and debtors were scrutinised on the risk cues used by the investigating auditor. Stocks and debtors were chosen because in these sub-accounts a relatively high error rate may be expected. And obviously in order to be able to validate risk assessment on (some quantity related to) the error rate, there has to be some variability in the error rate. Also the audit manual of the same firm which allowed us to use the (strictly anonymised) cases was used as a source of risk indicators. A third source of indicators was found in the audit manual of the audit departments of the Dutch ministries (HCDAD, 1997). The pilot resulted in a set of 13 risk indicators:

1. changes in the internal controls since last audit;
2. quality with which the administrative procedures are documented;
3. quality with which the authorisations are documented;
4. quality of separation of duties with respect to automation;

5. quality of the other separations of duties;
6. test and acceptance procedures of software;
7. security regarding access to EDP system;
8. operation of the other separations of duties;
9. violations of procedures and regulations;
10. violations of authorisations;
11. attitude of management towards internal control ;
12. expertise of personnel;
13. number and size of errors in previous audit.

The indicators from this pilot study were further developed by studying auditing literature with a view on key features for the internal control and for contextual factors (see for instance Arens & Loebbecke (Ch. 8,9, 1997)). This added 10 indicators:

1. strength of the system of controls;
2. access to other systems and assets;
3. are the processes a routine?;
4. the degree of pressure for high performance;
5. the occurrence of structural changes in the organisation;
6. the auditability of applicable law, regulations;
7. existence of audit trail;
8. the nature of the organisation;
9. the attitude of personnel;
10. the complexity of the organisation

So in this study 23 indicators were in use. Not every indicator was used with every organisation that was in the study; on the average 20 indicators per organisation were used. Indicator 11 (on manual intervention) in the questionnaire of this appendix was not present for all organisations; therefore it is missing in the analyses made for this thesis.

On the next pages the Questionnaire is given, its layout included.

QUESTIONNAIRE RISK ANALYSIS

RESEARCH PROJECT RISK ASSESSMENT

LIMPERG INSTITUUT

INTRODUCTION

Risk assessment is an essential part of auditing. Also with your organisation in many audits some form of risk analysis and assessment takes place. The consequences of risk analysis can be far reaching: they can affect both scope and depth of the audit activities that follow in the next phase. In spite of this key role, the quantitave aspects of risk assessment are still subject to serious discussions and even doubts. Trigger of this discussion is the transformation of qualitative (professional) judgements into quantitative assessments of chances that some event will occur and the inevitably subjective character of this transformation.

The questionnaire in front of you is meant to collect data in a research project that aims at forcing back this subjectivity. The key idea behind this is that by breaking down the risk assessment from two dimensions (inherent and control risk) into some 20 dimensions, the subjectivity in the judgement of how these dimensions interact, is reduced. In the questionnaire dimensions take the form of risk indicators, that you are asked to score, and questions about the organisation to which the financial statement applies that you have been auditing. In addition it asks for the size of the financial error you found in the audit. The aggregation of the separate risk assessments into some prior distribution on the error will be done on a statistical, empirical basis, as is explained in the next paragraph.

By gathering this type of data from a number of audits, say 30 as a minimum number, a data base is formed that gives the opportunity to analyse the relation between risk assessment and error found, on a purely statistical basis (given the data). This will lead to a calibration of the risk assessment of the respondents, and of the organisation they work for, to the extent that the respondents can be considered to be a random sample from the organisation. In a pilot study, conducted under the auspices of the Limperg Instituut, some results in this direction have already been gained. The ‘Limperg Instituut’ is an inter university institute in Amsterdam, for research into accounting and auditing. In the next phase the (regression like) relation between risk assessments and error will be used to predict the error rate that will be found in an new audit. This prediction may be given the form of a prior distribution of the error rate. An opening to a Bayesian approach with an empirically based prior distribution comes within reach in the near future.

In this questionnaire we ask for two types of answers:
- The general aspects you take into consideration when you assess the risk of a material error.
- Detailed data from a case from your own audit practice: risk assessment, size of the accounts, the error found

Filling out the questionnaire will take about one hour. The results of this research will be communicated to you (and possibly other people who are interested) in a presentation for you organisation. You will be invited in due time.
Whenever you have questions you can contact *the contact person* or Ed Broeze
We thank you in advance for your cooperation

Drs. Ed Broeze, *Contact person*

(adresses and telephone numbers)

I GENERAL FACTORS CONSIDERED TO BE RELEVANT FOR THE RISK OF A MATERIAL ERROR IN THE STATEMENT (THE RISK ASSESSMENT)

Please indicate by giving them a rank in the next table, which of the general factors mentioned you consider to be important, when assessing the risk of a material error (1 for the most important factor, etc).
If there are other factors that you think important, please add and rank them (cells 7, 8, 9, 10).

	Factor	**Ranking**
1.	The audit findings from audits of the same accounts of previous periods	
2.	Changes in the sector in which the organisation operates (new laws, concentration, heavy workload, etc.)	
3.	Design of the administrative system and controls	
4.	The extent to which data processing is automated	
5.	Attitude of management with respect to internal control	
6.	Major changes in the way the business operates	
7.		
8.		
9.		
10.		

II. YOUR AUDIT CASE

1. Characteristics of the audit object

(All amounts in thousands)

1.1

Which type of account or account area is the audit about? (multiple answers are possible)	article	
	Agency account	
	directorate account	
	resource account	
	any other	

1.2

What is the size in MU (*currency)* of the financial statement under audit?	KMU

1.3

What are the number of line items that make up this account?	

1.4

Could you estimate the percentage of largest items that make up 80% of the account?	%

1.5

Does the account area regard a flow of income or of expenditure?	In	out

1.7

Is the expenditure (multiple answers are possible)	subsidies	
	construction	
	cost of administration	
	contracts	
	others	

1.8

Has the administrative process been automated to a high extent?	Yes	no

1.9

Size of the (estimated) error last time	KMU	not available

2. Materiality

(Amounts in thousands of MU)

At what materiality limit did you assess the risks in the audit object	KMU = %

3. Probability of an error in the financial statement

3.1 General indicators

The table next page contains a number of risk indicators: (findings on) aspects that possibly have played a role in your assessment of the the occurrence risk (OR): the probability that (material) errors exist in the statement under audit.

In the column ‘Your conclusion’ you are asked to score every risk indicator according to your findings; either ‘-1’ when you consider the findings to be risk enhancing (e.g. many errors in the previous period), or ‘0’ when risk neutral , or ‘1’ when risk decreasing. If there are indicators that did not play a role in your risk assessment, you can show that in the rightmost column.

Indicator	Your conclusion (-1, 0 or 1)	Used in assessment? (yes/ no)
Relations with previous periods	XXXXX	XXXXXX
1 Number and size of errors in previous audits		
2 Changes in controls since last audit		
Design and existence of admin controls	XXXXX	XXXXXX
3 Strength of system of controls		
4 Quality of documentation		
5 Strength of segration of duties with respect to EDP		
6 Strength of segration of duties with respect to other procedures		
Working of admin procedures	XXXXX	XXXXXX
7 Acces to EDP systems (logical, physical)		
8 Access to other systems and assets (logical, physical)		
9 Extent to which authorisations and segregations are maintained		
10 Extent to which transactions can routinely be processed		
11 Extent to which there is manual intervention in data processing		
The context of the organisation	XXXXX	XXXXXX
12 Pressure for high performance		
13 Structural changes (mergers, privatisation, etc.)		
Auditability	XXXXX	XXXXXX
14 Auditability of law, regulations		
15 Existence of audit trail		
16 Nature of the organisation		
17Complexity of organisation structure		

Indicator	Your conclusion -1, 0 or1	Used in assessment? (yes / no)
General attitude towards internal control measures	XXXXX	XXXXXX
18 Attitude of management towards internal control		
19 Attitude of personnel towards internal control		
20 Professionalism/ expertise personnel		

3.2 Specific factors

Please mention (ranked in order of importance) specific factors concerning the relevant data processing that in your opinion **decrease** the probability that the financial statement will appear to be in material error (at most 3).

1.	
2.	
3.	

Please mention (ranked in order of importance) specific factors concerning the relevant data processing that in your opinion **increase** the probability that the financial statement will appear to be in material error (at most 3).

1.	
2.	
3.	

3.3 Costs of the risk assessment

Is it a new or a repeat audit?	new	repeat
How much audit effort did you need for your risk assessment?	.....days	
How many tests of the working of the system were part of that effort?	.....systems tests	

Please give your global risk assessment (of the occurrence risk)
in the following table (as you made it up before the substantive testing).

I deemed the occurrence risk to be

Very low	
Low	
Medium	
High	
Any other qualification (please describe)	

4. Audit design and -result

4.1

Do you distinguish large and small items?	Yes	no

4.2

If so, what is the boundary between 'small' and 'large''? (*in thousands)*	KMU

4.3

Design/ result	**large items**	**small items**
All items audited		
Judgmental sample		
Statistical line item sample		
Statistical monetary unit sample		
Something else, namely		
Sample size (also with judgmental sample		
Size of the error(s) found (*in MU)*		
Sum of the taintings		
Number of items in error in the sample		
Other relevant characteristics of the audit (please describe) *in a 'multiple sample situation' please give the data asked for on a separate account.*		

Summary

In **chapter 1** we relate how common practice in auditing allows a significant reduction of substantive audit when one of the outcomes of risk analysis, the occurrence risk, is assessed at the level 'low'. The occurrence risk is the risk (or probability) that the annual accounts, as they are presented to the auditor for his audit, contain a material error (an error of a size to be in the way of approval of the accounts). It is the result of the assessment of

- the inherent risk: the risk due to risk factors in the environment of the audit object and
- the control risk, the risk the accounting processes imply when dealing with the environmental risks and when safeguarding the proper operation of the business processes where they lead to recordings in the various accounts.

This practice of reduction of substantive work, guided by tables, raised questions. Can tables be justified, in use in the audit practice, that relate assessed occurrence risk to that decrease in substantive audit? Does this implicit quantification have an empirically sound basis? To find an answer to these questions, we engaged in a research into the validity of the assessment of occurrence risk. We hoped to find out how assessed occurrence risk relates to the 'real risk'. For this we had to get a better view on the 'real risk' and we needed validation criteria.

In **chapter 2** we develop a justification for four validation criteria:

1. the error rate and 3 measures constructed with the error rate as a basis.:
2. the 'audit position' of the error rate (is it larger than materiality?),
3. the 'sampling risk' (the risk a material error is not detected, due to sampling),
4. the empirical distribution of the error rates in the set of all cases (for a low occurrence risk the lower error rates should be prominent in this distribution, for a higher occurrence risk the higher error rates should be prominent).

The assessment of the occurrence risk aims at evaluating the quality of the administrative processes. For this quality the error rate is a direct indicator, but due to a random effect in the outcome, processes of the same quality will generate accounts with varying error rates. In combination with materiality (the limit for approval) the error rate gives a more conclusive indication of the risk that the administrative processes as they are analysed may produce a material error. This combination is made in the criteria "audit position" and "sampling risk". Because of this, we expect them to be better validation criteria. We were helped to justify the use of these criteria via the exploration of the Ensemble Prediction System (EPS) - the basis of weather forecasts - and by exploring the risk of water (the error) flooding dikes (the materiality); as metaphors. The exploration made clear that the empirical distribution of errors validates risk assessment.

In **chapter 3** we explore the logic of the Audit Risk Model (ARM) and give an overview of research. We explore the 'heuristics and biases' paradigm of Tversky and Kahneman (T&K). This paradigm is basic for dealing with risk assessment and for a majority of the available research. We explore the ARM, as to the event structure, the statistical validity and the level at which risk is assessed. We conclude that improper outcomes lie in wait, due to shortcomings with respect to the three aspects mentioned. Next to an ordering of the available research at the T&K paradigm, we use 'consistency', 'complexity' and 'operation of controls' as orientations in organising the available literature.
Our overview shows that

- the biases in the T&K paradigm do exist in risk assessment, but that with proper training they can be controlled
- consistency of risk assessment with relevant criteria exists; especially with respect to error rate. But unfortunately this is not stable over studies
- assessment at the level of assertions may improve risk assessment
- complexity of the audit object negatively affects assessments, but that experience improves handling of complexity
- the few studies on the effectiveness of tests of controls and on the influence of attributes of the administrative system indicate predictive power for the error rate, but in spite of that insufficient prevention of human errors

So there is much research on risk assessment, with positive outcomes. But the vast majority of this research applies only to aspects of this judgmental task. For a conclusive answer to the question, whether the current practice of quantification of risk is allowed, the assessed occurrence risk and (the four) validation criteria (our indicators for the 'real risk') have to be related. This will be the aim of our research.
Studies that indicate predictive power for distinct aspects also indicate that decomposition of risk assessment will improve its validity. This finding is the cause for our choice, in the first study, not only to validate classical risk assessment, but also to investigate possibilities for improvement by decomposing risk assessment into the assessment of risk indicators.

In **chapter 4**, as a preliminary, we solve the problem how to calculate the sampling risk, based on the data we have, by performing a simulation study to show that the use of a Beta distribution, based the estimate of the error rate and the corresponding sample size, is fit for the calculation of the sampling risk. We also found strong indications that this way of evaluating sample results may give valid and sharper confidence bounds for the error in the account under audit than the customary Stringer bound.

In **chapter 5** we explain the design of our research. It consisted of 3 studies.
Eight organisations participated in the first. They filled out questionnaires with which data were given for about 20 audit cases concerning 1996 or 1997, each as to their assessed occurrence risk and error found. The answers were based on the actual audit files, as much as possible. Next to that, scores were asked for the same audit cases on roughly 20 risk indicators.
The first study was followed by a second one: a replication of the validation of classical risk assessment, as it was done in one part of the first study. Here the audit cases concerned 2001.
Finally in a third study 'system testing' was investigated as part of the underpinning of risk assessment. The second and third study were partly guided by the availability of data.
In this chapter we explain our choice for a field study from the desire to validate risk assessment in the ultimate realistic and comprehensive setting. And that we rejected an experimental study, because we could not find ways to approximate risk as it emerges (or hides) in 'reality' closely enough.

In **chapter 6** we report the results of the first study with respect to the validity of classical risk assessment. They varied with respect to validation criterion.

1. With respect to the "audit position" for one organisation only, a significant negative correlation was found. Exactly only half of the relevant correlations had a '+'- sign, meaning that (the assessment of) occurrence risk tends to be higher when 'audit position' changes from 0 (not OK) to 1 (OK), where with valid assessment a negative sign is to be expected.
2. With respect to the error rate validity turned out to be satisfactory for the pooled organisations: the correlation at this level was .43 (p-value <.01). Still there was

a problem: the correlation coefficient varied considerably over organisations: from roughly 0 up to .72. So although on the average risk assessment appears to be valid with respect to error rate, for an organisation this gives no guarantee for the validity of its own risk assessment.

3. With respect to the sampling risk, validity could not be shown: for the pooled organisations a negative correlation was found (where a positive one indicates validity). Twenty cases turned out to suffer a serious threat from ineffectiveness (insufficient audit effort, if sample size would be exclusively based on the assessed risk) and 19 cases a threat from serious inefficiency (more audit effort than needed). For two of the distinct organisations it appeared that risk assessment is valid with respect to this criterion; but for two others the correlations were negative to a considerable extent.
4. With respect to the distribution of the error rates, we again found a satisfactory validity, for the pooled organisations. We had insufficient data for the distinct organisations to perform this analysis.

The evaluation of this outcome cannot be unambiguous: two validation criteria showed relatively positive validity, albeit not stable over organisations; two criteria did not show validity. Analyses for more or less constant levels of materiality did not improve this picture; neither did analyses with potential moderator variables.
This instability over the organisations and the failure of the validity to improve for logical moderator variables led to the conclusion that validity of risk assessment is not self-evident for an organisation. So unless it has established this validity, the organisation should assume that validity of risk assessment with respect to the error rate is absent. The strongest conclusion can be that in risk assessment the auditor actually predicts the error and not the risk that this error will exceed materiality. We planned to test this conclusion in a replication, the second study reported in this thesis.

In **chapter 7** we investigated whether risk indicators would improve validity of risk assessment with respect to the error rate, and preliminary questions. For this part of the first study we developed a set of about 20 risk indicators, from a pilot study, scrutiny of audit files, audit handbooks, textbooks and interviews with auditors. We analysed the indicators in this set by means of (multiple) correlations and regression.
The indicators turned out to be consistent with the assessed occurrence risk: both the bivariate correlations with all and the multiple correlation with a standard subset of the indicators were satisfactory.
The indicators were less consistent with the error rate. The bivariate correlations showed relatively many with the 'wrong' (+) sign (6 out of 19), although the sign test showed that the number of the right (-) signs for the correlations was still significant. Many of the scatterplots we produced confirmed that the bivariate relations of error rate and risk indicators were weak.
Only for two organisations a regression model that predicted the error rate could be formulated that was significant at the 5% level and had more than 20% explaining power. But these two models were very sensitive to the omission of one predictor: the explaining power vanished. In addition many of the predictors (indicators) got the 'wrong' sign, which implies that for that indicator you have to assume that the expected error decreases when (assessed) risk on that indicator increases. This illogical finding can be explained by the phenomenon of 'suppression'. It does not invalidate the regression model, but makes it hard to interpret. In conclusion, the indicators do not account for a convincing prediction of the error rate.
We tried to improve prediction by regression analysis on factorscores and on scales, both constructed from the indicator scores. But none led to a consistent improvement. Our conclusion was that we could not show improvement of the prediction of the error rate by the use of risk indicators

Because of many unanswered questions regarding the use of indicators, for a continuation of our study we expected more from an approach in which the promising relation of occurrence risk and error rate was further investigated.

In **chapter 8** we report the results of the replication of the first study into validity of the occurrence risk. We were allowed to use data from four governmental audit organisations concerning 2001. Unfortunately the relatively strong correlations of occurrence risk and error rate of the first study in our research, were not found in this replication. At the level of an organisation, both error rate, 'audit position' and sampling risk correlated at a near zero, or insignificant level with the occurrence risk. We had insufficient data for analysis of the conditional distributions.
In this replication we extended our research questions into the question whether the "error of last year" would have predictive power for the "error of this year". For two organisations we could calculate the relevant correlation. The predictive power was virtually absent in both.
Our conclusion for the organisation level had to be that in our replication we only found counter indications for the validity of risk assessment.
For the pooled organisations we could establish a moderate, positive correlation between error rate and occurrence risk and also the empirical distribution of the error rates showed validity, more with respect to the occurrence than with the size of errors.

In **chapter 9** we try to find out whether system testing can be used as an underpinning of risk assessment, as this is done in practice. This mostly regards the cases where the occurrence risk is assessed as 'low'. In such a case it is allowed to decrease the extent of substantive testing, on the condition that the risk assessment is underpinned by system tests. In our view this implies that system tests have to be predictive for the substantive error.
We analysed the relation between system tests and the error both at the transaction level and at the account level, in a very large set of dual purpose tests, divided over 5 years (1995-1999).
In all years at the transaction level the occurrence of a system error was predictive for the occurrence of a substantive error. But still only in some 4% of the cases where a system error occurred, also a substantive error was found.
In 4 out of 5 years (1996-1999), we found that at the account level, the occurrence of a system error is predictive for the occurrence of a substantive error. In 2 out of 5 years (1998, 1999) we found, also at the account level, that the occurrence of a system error predicts the size of the substantive error.
In this third study we also analysed the predictive power at the account level of the error "of last year" for the error "of this year". This power was only found for the error size of 1999. As regards the occurrence of substantive error: here we found three significant correlations. Only the relation between the years 1995 and 1996 was absent.
Our conclusions were that system testing is not self evident as underpinning, but that there are conditions under which it has this power.

In **chapter 10** we summarise our conclusions; they regard the validity of the assessment of occurrence risk in the classical sense and the possibilities to improve this assessment by means of decomposition with risk indicators. Next to that they regard the quality of system tests as predictor of the error rate and the "error of last year" as predictor for the "error of this year". We gave them already at the end of the summary of the corresponding chapter.
Most important conclusion is, that, u*nless an audit organisation has sufficient evidence on the validity of its risk assessment, it should assume that this risk assessment is not valid in the sense that it may replace substantive audit.*

We also gave suggestions for continuation of this research. In these suggestions we give preference to the possibility to engage in a development project in which the empirical distribution of the error rates of a ‘suitable set’ is taken as a point of departure. This distribution may serve as a prior distribution for the error rate in a new audit. Risk analysis comes into the picture when this prior distribution is made dependent on the assessed risk (both for the distribution in the ‘suitable set’ and for the prior for the new audit). When risk assessment has a high predictive power, the conditional distributions will differ substantially; when the discriminating power is absent, the conditional distributions will be the same. But still the empirical distribution may lead to a reduction in the extent of substantive testing, if the empirical distribution gives much weight to low error rates. In a rare case it may lead to an increase in substantive testing, if the empirical distribution gives much weight to the higher error rates.

Provided the prior is derived from the empirical distribution with caution for changes and new risks and taking into account the preventive influence of auditing, this approach is ready for use as soon as the ‘suitable set’ is defined and the empirical distribution of its errors established. Research may lead to the assessment of the predictive power of risk assessment for the empirical distribution and in that sense to validation of risk assessment and even to its calibration.

Next to this approach with priors from suitable sets, we recommend to adopt another way of modelling the prior information. This circumvents the problems with the Audit Risk Model and can be integrated in a natural way in the approach with the empirical distribution from suitable sets.

Samenvatting

“Valideren van risico-inschattingen in accountantscontrole”

In **hoofdstuk 1** geven we aan hoe in de gebruikelijke praktijk van auditing het toegestaan is de gegevensgerichte controle aanzienlijk te verminderen, als een van de uitkomsten van risico analyse, het bestaansrisico, op 'laag' wordt ingeschat. Het bestaansrisico is het risico (of de kans) dat in de jaarrekening, zoals deze wordt voorgelegd aan de auditor, zich een materiële fout bevindt (een fout die zo groot is, dat hij goedkeuren in de weg staat). Dit risico is het resultaat van het inschatten van:

- het inherente risico: het risico op een materiële fout tengevolge van omgevingsfactoren;
- het interne controle risico (control risk): het risico dat de administratieve processen onvoldoende mogelijke fouten tengevolge van omgevingsfactoren en als gevolg van eigen onvolkomenheden voorkomen, opmerken en / of corrigeren, zodat een materiële fout kan resulteren.

In deze praktijk worden standaard tabellen gebruikt, die een ingeschat bestaansrisico omzetten in een toegestane vermindering van de gegevensgerichte controle. Dit gebruik riep vragen bij ons op: hoe kunnen deze tabellen worden gerechtvaardigd, is er een gezonde empirische basis voor het omzetten van een ingeschat bestaansrisico in een vermindering van de gegevens gerichte controle? Om een antwoord te vinden op deze vragen besloten we een onderzoek op te zetten naar de validiteit van het inschatten van het bestaansrisico. In dit onderzoek hoopten we helderheid te krijgen over het verband tussen een ingeschat bestaansrisico en het "werkelijke risico". Hiertoe hebben we meer inzicht in het ‘werkelijke risico’ en valideringscriteria nodig.

In **hoofdstuk 2** ontwikkelen we een rechtvaardiging voor vier valideringscriteria:

1. het foutpercentage (omvang van de fout als deel van de omvang van de rekening) en drie grootheden die op basis hiervan geconstrueerd kunnen worden:
2. de "audit positie" van het foutpercentage (is deze groter dan de materialiteit?),
3. het steekproefrisico (de kans dat een materiële fout niet wordt ontdekt, omdat slechts een steekproef is getrokken),
4. de empirische verdeling van de foutpercentages in de verzameling van alle gevallen die in ons onderzoek geanalyseerd zijn (voor valide inschattingen verwacht je hierin voor een laag bestaansrisico dominantie van de lagere foutpercentages en voor een hoger bestaansrisico van de hogere foutpercentages).

De inschatting van het bestaansrisico richt zich op het vaststellen van de kwaliteit van de administratieve processen. Voor deze kwaliteit is het foutpercentage een directe indicator. In steekproeven zullen processen van dezelfde kwaliteit een variatie in de foutpercentages laten zien. In combinatie met de materialiteit (de goedkeurgrens) kan het foutpercentage een betere indicatie van het risico geven dat de administratieve processen een materiële fout zouden kunnen voortbrengen. "Audit positie" en steekproefrisico zijn mede op de materialiteit gebaseerd. Daarom zullen zij in principe betere valideringscriteria zijn. We werden geholpen in het vinden van rechtvaardiging voor onze valideringscriteria door het "Ensemble Prediction System" - de basis voor weervoorspellingen - te exploreren en dit, naast het risico dat water (de fout) een dijk (de materialiteit) overstroomt, als metafoor te gebruiken. Deze exploratie gaf ook grond aan het gebruik van de empirische verdeling van foutpercentages ter validering van het inschatten van risico's.

In **hoofdstuk 3** onderzoeken we de logische eigenschappen van het Audit Risk Model (ARM) en geven een overzicht van relevant onderzoek. We behandelen het "heuristics and biases"-paradigma van Tversky en Kahneman (T&K). Dit paradigma vormt een basis voor het beschouwen van de risico inschattingen en voor een meerderheid van de beschikbare onderzoeken. We exploreren het ARM op zijn gebeurtenissen-structuur, zijn statistische validiteit en op het niveau waarop het risico wordt ingeschat. We concluderen dat oneigenlijke uitkomsten op de loer liggen, als gevolg van tekortkomingen m.b.t. de drie genoemde aspecten.
Voor het ordenen van de beschikbare onderzoeken gebruiken we naast het T&K paradigma ook "consistency", "complexiteit" en "werking van het systeem" als ordenende begrippen. Onze overzicht laat zien dat

- de "biases" van het T&K paradigma zich inderdaad voordoen bij het inschatten van risico's, maar dat zij met een geschikte training beperkt blijven
- het inschatten van risico's consistent is met relevante criteria; speciaal m.b.t. het foutpercentage. Maar ongelukkigerwijs is deze consistentie niet stabiel over de onderzoeken
- het inschatten van risico's op het niveau van "controle beweringen" (ook wel "controle criteria") tot verbeteren van de inschatting kan leiden
- complexiteit van het controle object een negatieve invloed heeft op de kwaliteit van de inschatting, maar dat ervaring het omgaan met complexiteit verbetert
- de paar onderzoeken naar de effectiviteit van systeemtests en naar de invloed van kenmerken van het administratieve systeem op het foutpercentage, wijzen op een voorspellende kracht voor het foutpercentage, maar ook de hoogste kwaliteit is niet sterk genoeg om fouten te voorkomen.

Er is dus veel onderzoek op het terrein van risico inschattingen, met positieve uitkomsten. Maar het grootste deel van dit onderzoek heeft slechts betrekking op aspecten van deze beoordeling. Voor een sluitend antwoord op de vraag of de huidige praktijk van kwantificering van de risico is geoorloofd, zullen we het verband tussen het ingeschatte bestaansrisico en (de 4) valideringscriteria (onze indicatoren voor "het werkelijke risico") moeten onderzoeken. Dat zal het doel van ons onderzoek zijn.
Uit onderzoek dat wijst op een voorspellende kracht van specifieke aspecten van het administratieve processen kan worden geconcludeerd dat het decomponeren van risico-inschattingen de validiteit ervan kan verbeteren. Deze conclusie heeft ons ertoe gebracht om in de eerste studie niet alleen de klassieke risico-inschattingen te valideren, maar ook mogelijkheden te onderzoeken voor verbetering hiervan door de risico-inschatting op te delen over het inschatten van risico-indicatoren.

Vooraf aan het behandelen van ons eigenlijke onderzoek lossen we in **hoofdstuk 4** het probleem op, hoe het steekproefrisico te berekenen met behulp van de data die wij tot onze beschikking hadden: het geschatte foutpercentage en de bijbehorende steekproef omvang. In een simulatie studie laten zien dat het gebruik van de Betaverdeling, waarin deze twee grootheden worden gebruikt, geschikt is voor het berekenen van het steekproefrisico. Als een extra vonden we sterke indicaties dat deze manier van het evalueren van steekproef resultaten een betrouwbaarheidsgrens geeft voor de fout in de jaarrekening, die valide is en scherper dan de gebruikelijke Stringer bound.

In **hoofdstuk 5** verantwoorden wij onze onderzoeksopzet. Acht organisaties deden mee in de eerste studie. Hen werd een vragenlijst voorgelegd naar gegevens over ongeveer 20 controle gevallen (uit de jaren 1996, 1997): het foutpercentage het ingeschatte bestaansrisico en scores op de risico indicatoren. De antwoorden waren zoveel mogelijk gebaseerd op de controledossiers.

De eerste studie werd gevolgd door een tweede: deze hield de replicatie in van het valideren van het klassieke ingeschatte bestaansrisico, zoals dat al in de eerste studie was gedaan.
Tenslotte werd in een derde studie de voorspellende kracht van "systeemtests" onderzocht. De tweede en derde studie werden gedeeltelijk bepaald door de beschikbaarheid van de noodzakelijke data.
In hoofdstuk 5 lichten we ook onze keus voor een veldstudie toe. Hij is gebaseerd op de wens om risico inschattingen te valideren in realistische zettingen. We kozen niet voor een experimentele aanpak, omdat het ons onmogelijk lijkt risico's dicht genoeg te benaderen, zoals die blijken of verstopt zijn in de werkelijkheid.

In **hoofdstuk 6** rapporteren wij de resultaten van de eerste studie naar de validiteit van de klassieke risico inschatting. Zij varieerden per valideringscriterium, zoals de volgende samenvatting laat zien.

1. Met betrekking tot de "audit positie" werd slechts voor één organisatie een significante negatieve correlatie gevonden. Precies de helft van de relevante correlaties had een '+'- teken; in die gevallen bestond dus een tendentie dat een hoog ingeschat risico samenging met een 'audit positie' die goedkeuren niet in de weg stond. Bij een valide risico inschatting zou je juist het omgekeerde (en dus een negatieve correlatie) verwachten.
2. De validiteit met het foutpercentage als criterium was bevredigend voor de gepoolde organisaties: de relevante correlatie was hier .43 (p-waarde < .01). Toch was er een probleem: dezelfde correlatiecoëfficiënt per organisatie varieerde sterk: van 0 tot .72. Hoewel dus gemiddeld de risico inschatting valide is m.b.t. het foutpercentage, geeft dit aan een organisatie geen garantie m.b.t. zijn eigen risico-inschattingen.
3. De validiteit m.b.t. het steekproefrisico kan niet worden aangetoond: voor de gepoolde organisaties werd een negatieve correlatie gevonden (terwijl validiteit om een positieve vraagt). In twintig van de geanalyseerde audit gevallen bleek er een serieus gevaar voor te weinig controle inspanning te bestaan en in negentien gevallen bleek er een gevaar van een te hoge controle inspanning (meer dan nodig) te bestaan. In het eerste geval dus ineffectiviteit, in het tweede inefficiëntie. Geanalyseerd per organisatie bleek dat bij twee organisaties de risico-inschatting valide was op dit criterium (een sterke positieve correlatie); bij twee andere was de relevante correlatie juist sterk negatief.
4. We vonden een bevredigende mate van validiteit met betrekking tot de empirische verdeling van foutpercentages op het niveau van de gepoolde organisaties. We hadden niet genoeg data om deze analyse op het niveau van de afzonderlijke organisaties uit te voeren.

Deze uitkomsten zijn niet eenduidig: op 2 valideringscriteria vonden wij validiteit, ook al was deze niet stabiel over de organisaties; op 2 criteria vonden we geen validiteit. Analyses op een min of meer constant niveau van de materialiteit leidde niet tot verbetering van dit beeld, evenmin als analyses met potentiële moderator variabelen. Deze instabiliteit en het niet verbeteren voor potentiële moderator variabelen leidde tot de conclusie dat risico-inschatting door een organisatie niet vanzelfsprekend valide is. Tenzij hij deze validiteit heeft vastgesteld, moeten een organisatie er dus van uitgaan dat zijn risico-inschattingen niet valide zijn in de zin van ons onderzoek.
Onze bevindingen maken de conclusie aannemelijk dat een auditor eerder het foutpercentage dan het risico inschat. Met de bedoeling deze conclusie te testen zetten wij onze tweede studie op (zie hfdstk 8).

In **hoofdstuk 7** onderzoeken wij of het gebruik van risico-indicatoren de validiteit van risico-inschattingen met betrekking tot het foutpercentage verbetert. Voor dit deel van onze eerste studie ontwikkelden wij een verzameling van rond de 20 risico-indicatoren,

met behulp van een voorstudie, door het analyseren van controledossiers, controle-handboeken, tekstboeken en interviews met auditors. Wij analyseerden het verband tussen de indicatoren en het foutpercentage met behulp van (multipele) correlaties en regressie.
De indicatoren bleken consistent te zijn met het ingeschatte bestaansrisico: zowel de bivariate correlaties als de multiple correlaties, met een door de hele analyse gehanteerde standaard deelverzameling van indicatoren, waren bevredigend.
De indicatoren waren minder consistent met het foutpercentage. Van de bivariate correlaties hadden er veel het 'verkeerde' (+) teken (6 van de 19) ook al liet de tekentoets zien dat het aantal "juiste" (-) tekens significant was (op het 5% niveau). Vele spreidingsdiagrammen bevestigden dat de bivariate relatie van foutpercentage en indicatoren zwak is.
Slechts voor twee organisaties werd een regressiemodel gevonden dat het foutpercentage voorspelde en dat significant was op het 5% niveau en dat meer dan 20 procent verklaarde variantie (lees: "verklarende kracht") had. Maar deze twee modellen bleken erg gevoelig voor het wegnemen van een indicator: dan verdween de verklarende kracht vrijwel geheel. Bovendien hadden vele indicatoren het "verkeerde" teken, wat zou betekenen dat je voor zo'n indicator moet aannemen dat de verwachte fout afneemt als het op die indicator ingeschatte risico toeneemt. Die onlogische bevinding kan worden verklaard door het verschijnsel van "suppressie". Het regressie model is daarmee wel valide, maar het kan nauwelijks nog geïnterpreteerd worden. Al met al hebben wij met de indicatoren geen overtuigende predictoren voor het foutpercentage weten te vinden.
We hebben geprobeerd de voorspellende kracht te vergroten door regressie op factorscores en op door onszelf geconstrueerde schalen, beide gebaseerd op de indicatoren, te analyseren. Dit leidde niet tot een consistente verbetering. Onze conclusie kan alleen zijn dat wij geen verbetering van de voorspelling van het foutpercentage door het gebruik van risico-indicatoren konden aantonen.
Omdat er rondom de risico-indicatoren veel onbeantwoorde vragen waren, kozen wij voor een voortzetting van ons onderzoek waarin de meer belovende relatie tussen bestaansrisico en foutpercentage verder zou worden geëxploreerd.

In **hoofdstuk 8** rapporteren we de resultaten van deze gedeeltelijke replicatie van de eerste studie betreffende de validiteit van het bestaansrisico. Hierbij kregen wij toegang tot data in de controle dossiers over rekeningen uit 2001 van 4 departementale audit diensten. Helaas werden de relatief sterke correlaties van bestaansrisico en foutpercentage van de eerste studie, niet teruggevonden in deze replicatie.
Op organisatieniveau waren de correlaties van bestaansrisico met foutpercentage of steekproefrisico vrijwel gelijk aan 0. Ook de correlaties met "audit positie" waren niet significant. Voor analyses met de conditionele verdeling hadden we te weinig gevallen per organisatie.
In deze replicatie breidden we onze onderzoeksvragen uit tot de vraag of de "fout van het vorige jaar" de voorspellende kracht heeft voor de "fout van dit jaar". Voor twee organisaties konden we de benodigde correlaties uitrekenen. In beide gevallen waren deze vrijwel gelijk aan 0, dus de voorspellende kracht niet aantoonbaar.
Op organisatieniveau leidde dit tot de conclusie dat onze replicatie alleen aanwijzingen gaf voor het afwezig zijn van validiteit.
Op het niveau van de gepoolde organisaties vonden we een positieve correlatie tussen foutpercentage en bestaansrisico. Dit werd ook teruggevonden met de conditionele verdeling; deze liet zien dat het bestaansrisico eerder het voorkomen dan de omvang van een fout voorspelt.

In **hoofdstuk 9** onderzoeken we of systeemtests gebruikt kunnen worden als onderbouwing van de inschatting van het bestaansrisico, zoals dit in de praktijk gebruikelijk is. Deze onderbouwing betreft meestal de gevallen waarin het

bestaansrisico als "de laag" is ingeschat. In zo'n geval staat, zoals al gezegd, de audit methodologie toe de omvang van de gegevens gerichte controle te verminderen, mits de risico-inschatting wordt onderbouwd door systeemtests. Dit betekent, naar ons inzicht, dat systeemtests voorspellend moeten zijn voor de omvang van de fout.
Om dit te onderzoeken analyseerden wij de relatie tussen systeemtests en de fout zowel op het niveau van een transactie als op het niveau van een rekening, in een zeer grote verzameling van "dual purpose tests", verdeeld over vijf jaar (1995 tot en met 1999). Op het niveau van een transactie was in alle jaren het vóórkomen van een systeemfout voorspellend voor het vóórkomen van een gegevensfout. Dat nam niet weg dat slechts in ca. 4% van de transacties waarin systeemfouten voorkwam er ook een gegevensfout werd gevonden. Op het niveau van een rekening vonden wij in vier van de vijf jaren (1996 tot en met 1999) dat het voorkomen van een systeemfout voorspellend is voor het voorkomen van een gegevensfout. Voor de jaren '98 en '99 vonden we bovendien dat het voorkomen van een systeemfout voorspellend is voor de omvang van de gegevensfout.
In deze derde studie analyseerden wij ook de voorspellende kracht op het rekening niveau van de "fout van vorig jaar" voor de "fout van dit jaar". Voor de foutomvang werd deze alleen gevonden voor het jaar 1999, voor het voorkomen van een gegevensfout vonden we drie significante correlaties; alleen de relatie tussen de jaren '95 en '96 was afwezig.
Onze conclusie was dat systeemtests niet vanzelfsprekend zijn als onderbouwing, maar wel deze werking kunnen hebben.

In **hoofdstuk 10** vatten we onze conclusie samen; zij betreffen de validiteit van het inschatten van het bestaansrisico op de klassieke manier en de mogelijkheden dit te verbeteren via decompositie met behulp van risico indicatoren. Daarnaast betreffen zij de kwaliteit van systeemtests als voorspeller van het foutpercentage en van de "fout van vorig jaar" voor de "fout van dit jaar".
Belangrijkste conclusie is dat een organisatie alleen mag aannemen dat zijn risico-inschattingen valide zijn, als hij dit ook empirisch heeft vastgesteld. Voorlopig mogen we niet aannemen dat deze inschattingen het “werkelijke risico” geven.

In dit hoofdstuk geven wij ook suggesties voor voortzetting van dit onderzoek. Hierbij gaat onze voorkeur uit naar de mogelijkheid een ontwikkelingsproject op te zetten waarin de empirische verdeling van het foutpercentage, zoals gevonden in een "geschikte verzameling" (van audit gevallen), als uitgangspunt wordt genomen. Deze verdeling kan als voorverdeling voor het foutpercentage dienen in een nieuw audit geval. Ook het inschatten van het bestaansrisico komt in beeld, omdat deze voorverdeling afhankelijk gemaakt kan worden van het ingeschatte risico. Dit kan zowel voor de verdeling in de "geschikte verzameling", als in de voorverdeling voor de nieuw audit. Als risico-inschatting een voorspellende kracht heeft, zullen de conditionele verdelingen verschillen; als de voorspellende kracht er niet is, zullen de conditionele verdeling samenvallen. In beide gevallen kan de empirische verdeling, gebruikt als voorverdeling, echter toch leiden tot een vermindering van de omvang van de gegevensgerichte controle. Dat zal met name gebeuren als deze verdeling veel gewicht geeft aan de lage foutpercentages. In enkele gevallen zal hij ook tot een vermeerdering van de controle inspanningen kunnen leiden: als veel gewicht aan hogere foutpercentages wordt gegeven.
Op voorwaarde dat bij de afleiding van de voorverdeling uit de empirische verdeling marges worden ingebouwd voor veranderingen en nieuwe risico's en ook het preventieve karakter van een controle niet uit het oog wordt verloren, is deze benadering bruikbaar als een "geschikte verzameling" is gedefinieerd en daarin de empirische verdeling van de foutpercentages is vastgesteld. Het voortgaande onderzoek kan dan leiden tot het vaststellen van de voorspellende kracht van risico-

inschattingen voor de empirische verdeling. Zo wordt deze risico-inschatting gevalideerd en zelfs gecalibreerd, waarmee het het “werkelijke risico” aangeeft. Naast deze aanpak met voorverdelingen uit "geschikte verzamelingen", bevelen wij aan om een andere manier van modelleren van á priori informatie te gebruiken. Daarmee worden de problemen met statistische validiteit van het Audit Risk Model vermeden en de door ons voorgestelde modellering kan op een vanzelfsprekende manier geïntegreerd worden in de aanpak met de empirische verdeling van "geschikte verzamelingen".