Sample Size and Error Containment Judgments in Non-Statistical Audit Sampling - The Sensitivity of Auditors to a Revision of Professional Standards

Dissertation

zur Erlangung des akademischen Grades

"Doctor rerum politicarum" (Dr. rer. pol.)

Technische Universität Dortmund

Fakultät Wirtschaftswissenschaften

Lehrstuhl für Internationale Rechnungslegung und Wirtschaftsprüfung

Vorgelegt von

Daniel Baumeister

Dortmund

2019

Content

List of Figures

List of Tables

List of Abbreviations

a.a.O.	am angeführten Ort
Abs.	Absatz
ACL	Audit Command Language
AICPA	American Institute of Certified Public Accountants
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
AU Section	Auditing Section (AICPA)
BS	Berufssatzung
bspw.	beispielsweise
CI	Confidence Interval
d.h.	das heißt
DPR	Deutsche Prüfstelle für Rechnungslegung
e.g.	exempli gratia/for example
EDV	elektronische Datenverarbeitung
EPS	Entwurf eines Prüfungsstandards
ERP	Enterprise-Resource-Planning
et al.	et alii/und andere
F&A	Fragen & Antworten
f.	folgende
FA	Fachausschuss
FREP	Financial Reporting Enforcement Panel
GAAS	Generally Accepted Auditing Standards
GDPdU	Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen
ggf.	gegebenenfalls
H	Hypothesis
HFA	Hauptfachausschuss

HGB	Handelsgesetzbuch
Hrsg.	Herausgeber
i.d.R	in der Regel
i.e.	id est
i.e.S.	im engeren Sinne
i.H.	in Höhe
i.S.	im Sinne
i.S.d.	im Sinne der/des
IAASB	International Auditing and Assurance Standards Board
IDEA	Interactive Data Extraction and Analysis
IDW	Institut der Wirtschaftsprüfer in Deutschland e.V.
IFAC	International Federation of Accountants
IKS	Internes Kontrollsystem
ISA	International Standards on Auditing
ISAR	International Symposium on Audit Research
IT	Informationstechnik
Ltd.	Limited
Mio.	Millionen
MUS	Monetary Unit Sampling
n.F.	neue Fassung
No.	Number
n.s.	not significant
o.A.	ohne Angabe
o.g.	oben genannt
p. a.	per annum
PCAOB	Public Company Accounting Oversight Board

PIE	public interest entity
PS	Prüfungsstandard
p-value	probability value
Rev.	Revised
SAS	Statements on Auditing Standards
SEC	Securities and Exchange Commission
SOX	Sarbanes Oxley Act
St	Stellungnahme
std. dev.	standard deviation
T	Tausend
Tz.	Textziffer
U.S.	United States
VBA	Visual Basic for Applications
vgl.	vergleiche
vs.	versus
z.B.	zum Beispiel

1 Introduction

Throughout the past four decades audit sampling theory and practice have consistently undergone major changes, induced by both regulators and practitioners. However, the parties concerned often produced strongly counteracting outcomes between the application of audit sampling procedures as suggested by professional auditing standards and the actual execution in the field (PCAOB 2014; Krahel and Titera 2015). The subject is still of particular importance in academic literature as well as in professional debate, since none of the protagonists involved has managed to develop the 'silver bullet' that makes a uniform audit sampling approach applicable (Graham et al. 2018; Durney et al. 2014). This persistence is attributable to several factors, such as recent accounting scandals (Sibelman 2014) as well as technological and environmental influences that drive changes in accounting systems with which auditing standards and procedures must keep pace. Since SOX (the Sarbanes Oxley Act), a growing demand for information by financial statement addressees has been further reflected in increased disclosure to the public (e.g. key audit matters).

In response to these informational demands, professional standards are becoming more extensive to cover detailed statements on the multifaceted execution of sampling procedures (AICPA 2017; IDW 2016c; Hitzig 2004). At the same time, such bulkiness builds up the potential of overstraining the auditor's cognitive capacity and might create a diametrically opposite transition of sampling practice (Fay et al. 2015; Trompeter and Wright 2010; PCAOB 2008). A proper execution of the sampling process by the auditor leads favorably to conclusions about the probability that an accounting population (e.g. an accounts receivable balance) has been misstated. However, a mathematical determination of the relevant risk components can be achieved only when statistical sampling is performed. While these procedures were predominant in auditing practice for a long time (e.g. Mochty 2012; PCAOB 2008), non-statistical sampling has become common nowadays (Christensen et al. 2015; Elder et al. 2013). However, the subjective assessment of input parameters makes non-statistical sampling prone to decision biases (AICPA 2017; Hall et al. 2012; IAASB 2009b; Swearingen and Hansen 1990).

For both techniques, professional standards set by the American Institute of Certified Public Accountants (AICPA), the International Auditing and Assurance Standards Board (IAASB) and the Public Company Accounting Oversight Board (PCAOB) require sufficient and appropriate audit evidence (AICPA 2011; IAASB 2009b; PCAOB 2014). While sufficiency is determined by the extent of substantive testing, appropriateness addresses

the relevance and reliability of the evaluation of disclosed misstatements. The significance of these dimensions is accompanied by a recent revision of German professional audit sampling standards (IDW 2016a, 2016b). In particular, the introduction of ISA 530 'audit sampling'-based content into the revised IDW PS 300 n.F. and IDW PS 310 supersedes the formerly effective IDW St/HFA 1/1988. With the latter having focused on statistical rather than non-statistical procedures, the new standards will become mandatory in German audits from 2018. Consequently, not much is known about the effectiveness of audit sampling in the profession.

With Germany being the only country in which the observed change in auditor behavior is accompanied by a revision of professional standards, this provides a unique setting for research in audit sampling. Echoing the interest of researchers and practitioners in auditor behavior, this dissertation is aimed at investigating auditors' performance in crucial subprocesses of non-statistical audit sampling under these revisions. Thus, this research combines the behavioral research track with the effect of a recently changed audit standard. In this regard, previous audit sampling research focused mainly on methodological improvements rather than observing the effects of the introduction of new concepts on behavior (e.g. Zaheen et al. 2013; Swearingen and Hansen 1990; Leslie et al. 1979) and on performance (e.g. Hall et al. 2012; Hitzig 2004; Messier et al. 2001). The considerable rise in data-induced complexity in accounting systems has led to a growing body of research investigating the determinants and effects of tools aimed at mitigating the negative consequences of auditors' decisions in very broad settings. Observing and developing so-called decision aids (Bonner et al. 1996; Messier 1995), these studies focus on the reduction of cognitive strain through a decomposition of problems in auditing (e.g. Graham et al. 2018; Fay et al. 2015), as well as on the consequences of the changing nature of accounting misstatements (Griffin and Wright 2015; Krahel and Titera 2015; Durney et al. 2014). However, regarding audit sampling, various questions remain, on which this dissertation seeks to shed some light. In particular, the individual papers comprise conceptual frameworks (Papers 1, 2 and 4) and an archival study (Paper 3) on the status-quo of audit sampling practice as well as potential effects of the standards' update. Building on the archival results, two behavioral experiments were conducted (Papers 5 and 6) which focus on the execution of substantive testing procedures and address the question of whether extended guidance necessarily leads to more consistent and accurate decisions. Detecting these effects is important, as previous research has shown that changes in

auditing standards do not necessarily increase audit quality (Freitas et al. 2004; Messier et al. 2001; Burgstahler et al. 2000; Wheeler et al. 1997). With regard to audit sampling, prior research in this context has shown that auditors might be generally aware of crucial audit sampling properties, but deliberately circumvent normative requirements (Messier et al. 2001; Hermanson 1997; Kachelmeier and Messier 1990).

The *first paper* of this dissertation ('Stichproben in der Jahresabschlussprüfung – Voraussetzungen zur Anwendung anerkannter Erhebungsverfahren') provides insights into audit sampling guidance covered by the former IDW St/HFA 1/1988 and the changes anticipated by its update. A review of the technological and normative development of audit sampling exposes the demand for research to focus more on non-statistical than on statistical sampling plans. We analyze changes in the auditing environment towards the most disruptive problems regarding audit sampling. Our results show that sampling must be considered in any financial statement audit, regardless of a client's legal structure or of the industry and the present level of audit risk. While revealing the availability of multiple methodological options which facilitate a wide-ranging application of audit sampling, the paper highlights the fact that the actual implementation of any of these procedures is susceptible to procedural misconceptions. As auditing standards lack a uniform sampling concept, auditors are prone to neglecting error-specific properties. Because most statistical concepts are not suited to the statistical properties of accounting populations, many of the endorsed procedures put more obstacles in the auditor's way than solve the problem of selection.

Although several studies have analyzed bias-avoidance techniques (Fay et al. 2015; Messier et al. 2001; Leslie et al. 1979), research on the patterns of auditor behavior in sampling and the quality of their judgment is scarce. As subsequent reviews of sampling plans cannot identify decision failures owing to the unavailability of the true error value of an accounting population, it remains an open question as to what extent auditors reach faulty decisions. The *second paper*, ('Die praktische Durchführung von Auswahlprüfungen – Strategien zur Vermeidung wesentlicher Fallstricke bei der Erhebung und Auswertung von Stichproben') provides a conceptual framework and aims to design strategies in the avoidance of pitfalls identified in the first paper by screening decision opportunities within crucial sampling plan subprocesses. Our study brings evidence to this issue by narrowing down the most likely occurrences of misconduct and debating how to manage them by adjusting auditor behavior. In line with prior research from the

U.S., we conclude that erroneous auditor performance can be largely attributed to an inadequate determination of the scope of an audit (Allen and Elder 2005; Messier et al. 2001) and a faulty evaluation of accounting errors revealed during the audit (Fay et al. 2015; Mauldin and Wolfe 2014; Messier et al. 2001).

At this stage, the investigation is still abstracted from the tenor of professional standards. Certainly, the search for improvement must be relevant to auditors' field work (AICPA 2017; Griffin and Wright 2015; Jacoby and Hitzig 2011). For example, using a bound estimate to test inventory accounts is reasonable only if the distribution of errors within the underlying population can be assumed to be normal (Peek et al. 1991; Johnson et al. 1981). However, the statistical characteristics of accounting populations in the field are not comparable with those that are usually considered in research in the field of statistics (Giezek 2014; Zaheen et al. 2013; Fienberg et al. 1977). Errors in accounting populations can be assumed to be rare, ranging around less than .01 of individual book entries (Durney et al. 2014; Arens and Loebbecke 1999). Additionally, causes of errors can vary largely between simple transcription mistakes and fraudulent management override; hence, there often is no tangible 'right or wrong'. Considering this, the auditor must recognize a certain qualitative dimension to misstatements, even when the actual choice of sampling technique is correct.

As IDW St/HFA 1/1988 provides only relatively general indications for the composition of the sampling process, more detailed guidance such as the AICPA audit guide 'Audit Sampling' (AICPA 2017) may be worthwhile for national purposes. Hence, the *third paper* ('Prüfungsnachweise mit Hilfe von Auswahlprüfungen – Empirische Analyse unter Berücksichtigung der Änderungen durch IDW PS 300 n.F. und IDW PS 310') is an archival study investigating the execution of audit sampling in the field before the new standards come into effect. We examine working papers from 340 individual sampling applications in various auditing settings, set among highly diverse financial statement audits, and taken throughout audit engagements in 73 small private to large public enterprises in the fiscal year 2017. We initially investigate the responses of professional auditors to client-specific properties as well as the characteristics of errors and individual risk functions attributed to the engagements. Our analysis focuses on the consistency of these factors with the choice of audit sampling procedures, auditing scope, and the handling of accounting errors. Engagements to be considered in the study were selected with the particular aim of reflecting the typical client structure of a large second-tier auditing firm.

Hence, we were able to draw conclusions about audits of small and medium-sized companies as well as public interest entities (PIEs), with the latter being exposed to comprehensive reviews by external oversight institutions such as the U.S. Securities and Exchange Commission (SEC) and the Financial Reporting Enforcement Panel (FREP/Deutsche Prüfstelle für Rechnungslegung (DPR)). We find that non-statistical sampling is the most common sampling procedure, used for almost every auditing objective. Flaws can be observed regarding the determination of a reasonable sample size and the accurate treatment of 'uncommon' errors. Tracking down the apparent difficulties, we show that inherent and control risk do not have a significant impact on audit scope and error projection. Moreover, our findings demonstrate that experience in auditing and the availability of working papers from the previous year play a significant role in the conversion of error-specific information into sampling properties.

At the time the study was conducted, drafts of the revised standards were being debated by scientists and practitioners. Regarding the work-over, our results imply a correlation between the new guidance and the elaboration of audit sampling. Neither of the key factors of sample size and error projection is sufficiently associated with perceptions of similarity among auditors, for several possible reasons. For example, error projection might be circumvented because of the necessity of expanding a sample, which would eventually lead to disconcerting discussions with clients. With respect to knowledge, an inexperienced auditor might not have encountered a wide variety of error causes and label more errors as 'unique'. A revision of professional standards potentially has an indicatory effect on these issues. Since paper 1 shows that the update might not be a stand-alone solution, the question arises as to whether ISA 530 itself lacks sufficient detail in crucial components of the sampling process. Thus, the *fourth paper* ('Prüferisches Ermessen vs. mathematische Präzision – Schränkt IDW PS 310 die Handlungsfreiheit des Abschlussprüfers ein?') focuses on the potential impact of the forthcoming revision on auditor behavior. Combined with the results from the archival study (Paper 3), this demonstrates the possibility of a positive correlation between the new standards and the mitigation of identified flaws. While some prior studies have documented a positive effect of decision aids on auditor behavior (Bonner et al. 1996; Ashton 1990; Akresh and Tatum 1988), little is known about the potential of implementing decision aids into auditing standards when the preceding guidance was deliberately circumvented (Trompeter and Wright 2010; Messier et al. 2001; Kachelmeier and Messier 1990).

To test for these assumptions two experiments were conducted. Given opposite views on the accuracy of auditor behavior regarding audit scope, the *fifth paper* ('Does (Sample) Size Matter? The Sensitivity of Auditors to a Revision of Non-Statistical Audit Sampling Standards') addresses the impact of an ISA 530-based semi-structured decision aid on sample size appraisals among professional auditors. Setting sample size based on one's judgment is a complex task, as multiple conflicting parameters need to be considered. Research in auditing (Messier et al. 2001; Elder and Allen 2003) and psychology (Tversky and Kahneman 1971, 1974; Kahneman and Tversky 1972) has indicated that auditors utilize heuristics which simplify the profound judgmental operations but also increase the risk of misallocation. As the revised standards include qualitative guidance on the judgmental determination of sample size, we conducted a 2 x 2 x 2 between-subjects experiment, in which we manipulated the presence of guidance as well as different levels of inherent and control risk and the number of additionally performed audit procedures. The two latter factors are explicitly included in the standards' decision aid. Drawing on anchoring and adjustment theory (Libby 1985; Tversky and Kahneman 1974), we tested our hypotheses among 179 professional auditors employed in various auditing firms. We found that auditors are generally aware of crucial parameters affecting sample size. However, they face difficulties implementing this knowledge in the presence of unusual audit objectives, as they tend to lose sight of statistical concepts and excessively reduce the proposed sample size. For example, a majority of the participants tended to increase sample size disproportionately in high-risk audit settings. Consequently, audit risk became more acute than statistical concepts would permit. However, we find the decision aid to have a positive effect in common audit settings, rendering significantly larger samples. With respect to auditor litigation risk, participants performed poorly when compared to statistically derived samples. Showing that the decision aid makes a significant impact on participants' judgments but does not improve auditor consensus on accurate sample sizes, the results hold for two levels of client risk as well as for two levels of other substantive testing procedures that were performed, verifying that decision biases can be mitigated when a cognitive strain is moderated (e.g. Messier et al. 2001; Bonner et al. 1996). Overall, these findings have important implications for standard setters in that they draw attention to the frequent failure of non-statistical sampling plans to appropriately consider inherent and control risk, revealing a weakness in a majority of critical auditing situations. These implications should also be of interest to the policy departments

of auditing firms wishing to design sampling plans to prevent any circumvention of statistical concepts.

After audit scope, error evaluation is the factor most responsible for driving auditors' faulty judgment. Like statistical sampling, non-statistical procedures require the projection of revealed misstatements onto the underlying accounting population (IDW 2016b; IAASB 2009b). From the auditor's perspective, non-projection is a favorable option in resolving a sampling plan, as it prevents discussion with the client due to an expansion of sampling procedures as well as corrective book entries based on the auditors' estimate of total error. Studies have shown that auditors have difficulty applying the statistical principle of extrapolation and tend to label errors as unique, which leads to a non-projection of representative misstatements (Burgstahler and Jiambalvo 1986; Dusenbury et al. 1994; Burgstahler et al. 2000). In this context, the revised standards provide ISA 530-based qualitative guidance on the evaluation and resolution of audit samples, explicitly considering the possibility of errors being labeled as anomalies (IAASB 2009b; IDW 2016b). By contrast, PCAOB auditing standards and the AICPA audit sampling guide indicate the necessity for error projection in any case (PCAOB 2014; AICPA 2017). Even so, error containment is observed to be common practice, even beyond ISA-regulated audit settings (see also section 4; Christensen et al. 2015; Fay et al. 2015; Durney et al. 2014), driving auditors to treat qualitatively different errors in quantitatively different ways.

With these amendments potentially simplifying error isolation, the *sixth paper* ('How Many Needles Are in the Haystack? Sensitivity of Error Containment Judgments to Changes in Audit Standards') extends the view from the preceding experiment by addressing auditors' tendency not to project but rather contain errors from the remainder of an audited population (Durney et al. 2014; Elder et al. 2013). Besides providing a normative model of the risks associated with excluding errors, we examine the effect of error containment guidance on participants' perceived occurrence of anomalies in audit sampling. Drawing on the representative heuristic, we assume that the guidance reduces the tendency to remove misstatements from estimates of total error. We test our research question among 80 senior-level auditing students using a 2 x 2 x 2 mixed between- and within-subjects factorial design, in which we manipulate the extent of ISA 530 guidance, two levels of containment information and two levels of error frequency. All participants assessed the necessity of error projection for eight different errors raining from common to fraudulent errors revealed within the scope of a prototypical non-statistical sampling

plan. Contrary to the normative intent of ISA 530, the additional guidance does not significantly affect the participants' behavior and hence does not represent a reasonable bias avoidance technique. In fact, the results indicate that containment decisions are influenced mainly by the perceived error frequency, rather than the actual causes of errors. In line with prior research in auditing (Allen and Elder 2005; Hitzig 2001; Burgstahler et al. 2000; Hermanson 1997; Dusenbury et al. 1994; Burgstahler and Jiambalvo 1986) and psychology (Kahneman and Tversky 1972), we conclude that decision biases are present especially when encountered misstatements are resolved qualitatively. Additional guidance must therefore be much more specific than that encompassed in ISA 530. At the same time, our findings might be a further warning to standard setters and auditing firms, since additional guidance on error evaluation and error containment does not prevent the feared naturalness of containing representative errors.

To summarize, the present dissertation highlights the ongoing importance and potential positive effects of more versatile audit sampling procedures, but also the negative consequences of non-statistical sampling procedures and the necessity for a corresponding revision of professional standards. Any practicing auditor or researcher addressing audit quality must be aware of the crucial impact substantive testing procedures have on the merit of the auditor's opinion. This thesis indicates that audit sampling must still be considered a major and powerful tool in any audit of financial statements. However, the observed weaknesses of auditors' performance when applying non-statistical sampling point to a clear potential for improvement. It also sheds light on the necessity for audit regulators and auditing firm policy departments to keep pace with changing accounting environments, as the presence of large amounts of data not only changes the nature of errors, but also increases the impact that individual errors have on an estimate of total error.

This dissertation is a cumulative work consisting of six individual papers, of which two are working papers. Please note that some papers have already been published or will soon be under review for publication. Consequently, further modifications of individual paper versions presented in this dissertation will follow. Further versions of the papers will be available in the respective journals or at scientific platforms after publication. Thus, please be sure to cite only the latest versions of these papers.

2 Stichproben in der Jahresabschlussprüfung – Voraussetzungen zur Anwendung anerkannter Erhebungsverfahren

2.1 Publikationsdetails

Zusammenfassung: Bei der Planung und Durchführung von Funktions- und Einzelfallprüfungen hat der Abschlussprüfer gemäß IDW EPS 300 n.F. wirksame Verfahren zur Auswahl der zu prüfenden Elemente festzulegen, um ausreichende und angemessene Prüfungsnachweise zu erlangen. Dies wird regelmäßig durch eine Prüfung in Stichproben erreicht. Auswahlprüfungen zur Steigerung der Prüfungseffizienz bieten – eine methodisch einwandfreie Anwendung vorausgesetzt – eine überzeugende Bandbreite an Vorteilen, denen jedoch im Falle irrtümlicher Annahmen bei der Planung und Durchführung exorbitante Risiken gegenüberstehen. Ziel dieses Beitrags ist es daher das relevante Methodenspektrum vorzustellen und wesentliche Anwendungsvoraussetzungen der Durchführung von Auswahlprüfungen zu erörtern.

Koautoren: Christoph Oldewurtel.

Stichwörter: Prüfungsnachweise, Statistische Stichproben, Einzelfallprüfungshandlungen, IDW EPS 310, ISA 530.

Publikationsstatus: Veröffentlicht in: *WP Praxis* 5 (7): 169–175.

2.2 Bedeutung von Auswahlprüfungen für die Jahresabschlussprüfung

Ein methodisch und regulatorisch häufig diskutierter Themenkomplex in der Prüfungsliteratur ist die Verwendung von Stichproben im Rahmen der Jahresabschlussprüfung. Letztere stellt i.d.R. keine Vollprüfung sämtlicher Transaktionen eines Geschäftsjahres dar[1]. Folglich findet sich im Bestätigungsvermerk des Abschlussprüfers regelmäßig der Zusatz, dass Wertansätze und Nachweise für die Angaben in Buchführung, Jahresabschluss und Lagebericht (im Wesentlichen) auf der Basis von Stichproben beurteilt wurden. Dabei identifiziert der Prüfer wesentliche Prüffelder und würdigt diese auf Basis bekannter und während der Prüfung gewonnener Vorinformationen in einem zu spezifizierenden Umfang. Er bedient sich dieser Methodik insbesondere im Rahmen von Funktionsprüfungen des internen Kontrollsystems sowie aussagebezogenen Einzelfallprüfungen[2]. Zu diesem Zweck werden ausgewählte Posten (Stichprobe) eines interessierenden Prüffelds (Grundgesamtheit) genauer untersucht, um Aussagen über das gesamte Prüffeld machen zu können. Von zentralem Interesse sind vor allem die Häufigkeit sowie der Umfang von Fehlern im Prüffeld, wobei die Aussagesicherheit des Prüfungsurteils einer Unsicherheit (Stichprobenrisiko) unterliegt. Insbesondere das angewendete Auswahlverfahren sowie die Methode der Hochrechnung des Ergebnisses auf die Grundgesamtheit (Repräsentationsschluss) sind dabei auch mittels EDV-Unterstützung nicht ohne Weiteres problemlos zu beherrschen.

Im Gegensatz zur deutschen Prüfungsliteratur findet sich im anglo-amerikanischen Schrifttum eine schier unüberschaubare Anzahl publizierter Methoden und Anwendungshinweise[3]. In jüngerer Vergangenheit stand dabei die Fortentwicklung des klassischen sog. Monetary Unit Sampling (MUS) und weiterer Schätzverfahren im Mittelpunkt[4], wobei eine Adaption für die Praxis bis dato ausgeblieben ist[5]. Im Gegensatz zu den Veröffentlichungen des AICPA sind Hilfestellungen der Standardsetter auf nationaler und europäischer Ebene rar. Als Novelle in diesem Bereich kann jedoch die Ersetzung der bisher gültigen IDW St/HFA 1/1988 durch IDW EPS 310 als Umsetzung von ISA 530 gesehen

[1] Vgl. IDW PS 200, Tz. 19.
[2] Vgl. IDW (Hrsg.), WP-Handbuch, Band I, Düsseldorf 2012, R, Tz. 119, S. 2432.
[3] Vgl. z.B. AICPA (Hrsg.), Audit Guide – Audit Sampling, New York 2014; *Stewart*, Technical Notes on the AICPA Audit Guide Audit Sampling, New York 2012.
[4] Vgl. exemplarisch *Zaheen/Shabbir/Gupta*, Commun Stat Theor M 18/2013, S. 413-422; *Hoogduin/Hall/Tsay*, Auditing-J Pract Th 1/2010, S. 125-148.
[5] Vgl. *Giezek*, WPg 11/2014, S. 565; *Ruhnke/von Torklus*, WPg 23/2008, S. 1120. Demnach werden 85 % der Stichproben in der Prüfungspraxis auf Basis nichtstatistischer Verfahren oder bewusst erhoben.

werden[6]. Damit werden die Anforderungen an die Anwendung von Stichprobenverfahren in Prüfungen, die nicht bereits freiwillig den Verlautbarungen des IFAC entsprechen[7], quasi verbindlich. Als unmittelbare wesentliche Änderung ergibt sich hierbei, dass auf einer bewussten Auswahl basierende Methoden nicht (mehr) als Stichprobenverfahren deklariert werden. Können die Verfahren der bewussten Auswahl, der Zufallsauswahl sowie das MUS inzwischen als in der Praxis bekannt und etabliert angesehen werden, stellt sich folglich die Frage nach der Tragweite etwaiger Auswirkungen auf die Prüfungspraxis.

2.3 Motivation von Auswahlprüfungen

2.3.1 Erreichung von Prüfungseffizienz

Der Prüfer wählt zur Maximierung der Prüfungseffizienz und schlussendlich des Deckungsbeitrags[8] stets diejenige Handlungsalternative, mit der ein hinreichend sicheres Prüfungsurteil mit dem vermutlich geringsten Mitteleinsatz abgegeben werden kann[9]. Eine Aussage über die inhaltlich und wertmäßig korrekte Abbildung einzelner Transaktionen erlauben die Prüfung des internen Kontrollsystems sowie analytische Prüfungshandlungen dabei nur eingeschränkt, sodass i.d.R. Einzelfallprüfungen durchzuführen sein werden[10]. Auswahlprüfungen als Alternative zur Vollprüfung[11] kommen dabei insbesondere dann zum Tragen, wenn die Anzahl der Elemente im Prüffeld groß ist[12], wenige oder keine Informationen zu einem erhöhten Risiko spezifischer Elemente vorliegen und die technischen Voraussetzungen zur Datenerhebung seitens des Mandanten gegeben sind.

Als angestrebtes allgemeines Ziel kann dabei der Versuch angesehen werden, die Prüfungseffizienz durch Integration statistischer Gesetzmäßigkeiten in den Prüfungsansatz

[6] Im Rahmen des Clarity Project wurde die Darstellung der verfügbaren Prüfungsmethoden in ISA 500 verlagert. Dieser Struktur folgen auch die kommenden deutschen Prüfungsstandards.
[7] Vgl. dazu WPK (Hrsg.), Zeitpunkt der verpflichtenden Anwendung der ISA in Deutschland vom 07. 05. 2015, http://www.wpk.de/neu-auf-wpkde/alle/2015/sv/zeitpunkt-der-verpflichtenden-anwendung-der-isa-in-deutschland/, (Abruf am 10. 12. 2015).
[8] Vgl. IDW PS 200, Tz. 21, wonach Wirtschaftlichkeit zur Nebenbedingung der Prüfungsziele avanciert.
[9] Vgl. *von Wysocki*, Prüfungstheorie, messtheoretischer Ansatz, in: *Ballwieser/Coenenberg/von Wysocki (Hrsg.)*, Handwörterbuch der Rechnungslegung und Prüfung, 3. Auflage, Stuttgart 2002, Sp. 1707-1715.
[10] Vgl. IDW PS 261 n.F., Tz. 83.
[11] Vgl. IDW EPS 310, Tz. A49.
[12] Vgl. *Griffin/Wright*, Account Horiz 2/2015, S. 377-379 zum Einfluss immer umfangreicherer Rechnungslegungssysteme auf die Entwicklung der Jahresabschlussprüfung.

zu steigern[13]. Im Prüffeld wird dabei eine abgegrenzte Menge von Daten als statistische Masse oder Grundgesamtheit bezeichnet[14]. Eine Grundgesamtheit ist eine Menge von gleichartigen Elementen, z.B. sämtliche Kreditorensalden oder sämtliche zu einem Kreditorenkonto zugehörigen Rechnungen, wobei eine Stichprobe sodann die Teilmenge einer solchen Grundgesamtheit darstellt. Jedes Element der Stichprobe muss dabei genau einem Element der Grundgesamtheit entsprechen. Eine Stichprobe selbst wird nach bestimmten Kriterien ausgewählt und soll repräsentativ für die Grundgesamtheit sein, d.h. deren Eigenschaften widerspiegeln. Der Prüfer akzeptiert im Rahmen des risikoorientierten Prüfungsansatzes, dass nicht alle Elemente eines Prüffeldes geprüft werden und damit eine gewisse Toleranz bei seinen Prüfungsaussagen, das sog. Stichprobenrisiko. Eine Hochrechnung im statistischen Sinne unterliegt dabei engen Restriktionen, denen gemäß IDW EPS 310, Tz. 7 g) im Wesentlichen zwei Faktoren innewohnen:

- Zufallsgesteuerte Auswahl der zu prüfenden Elemente[15] sowie
- Anwendung der Wahrscheinlichkeitstheorie zur Auswertung der Stichprobenergebnisse, einschließlich der Bewertung des Stichprobenrisikos.

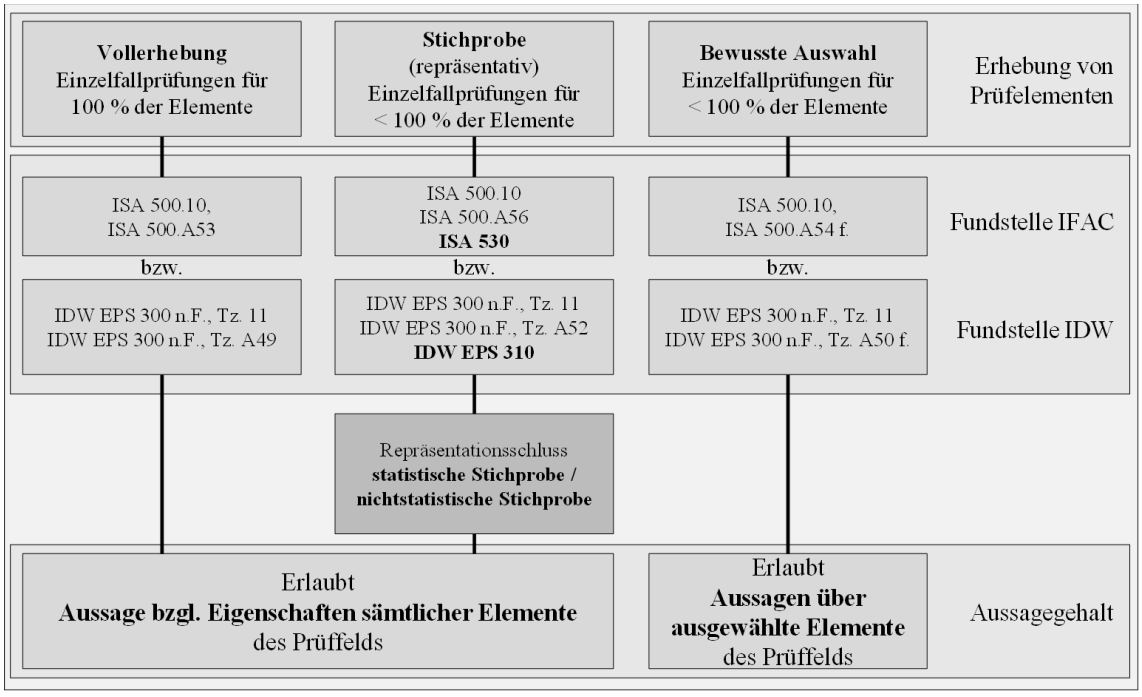


Abbildung 2.1: Übersicht zur Verfügung stehender Auswahlverfahren nach IFAC und IDW[16]

[13] Vgl. z.B. *Hitzig*, The CPA Journal 5/2004, S. 30.
[14] Vgl. *Guy/Carmichael/Whittington*, Audit Sampling, New York 2002, S. 2.
[15] Vgl. *Weinand/Wolz*, WP Praxis 4/2012 S. 69-73 für eine mathematische Darstellung der Techniken der Zufallsauswahl.
[16] Eigene Darstellung.

Erfüllt das zur Wahl stehende Verfahren nicht kumulativ beide Bedingungen, handelt es sich nicht um ein statistisches Stichprobenverfahren (vgl. auch Tabelle 2.1).

Unabhängig von der gewählten Auswahlmethode sollte sich der Prüfer stets darüber im Klaren darüber sein, dass eine rein quantitative Argumentation von Prüfungsnachweisen nicht hinreichend ist. IDW EPS 300 n.F., Tz. A5 stellt dahingehend klar, dass eine Kompensation von Prüfungsnachweisen von schlechter Qualität durch die Einholung zusätzlicher Prüfungsnachweise nicht möglich ist[17]. Zudem wird erst durch die Kombination der Ergebnisse stichprobengestützter Prüfungshandlungen mit weiteren aussagebezogenen Prüfungshandlungen ein Gesamturteil auf Basis eines Vergleichs der festgelegten Wesentlichkeitsgrenze bzw. tolerierbaren Abweichung mit dem hochgerechneten oder nach prüferischem Ermessen einzuschätzenden Gesamtfehler im Prüffeld möglich.

Tabelle 2.1: Abgrenzung statistischer/nichtstatistischer Stichprobenverfahren[18]

	Auswahlmethode	**Hochrechnung**
Statistische Stichprobe	zufällig	Mathematisch unter Berücksichtigung des Stichprobenrisikos
Nichtstatistische Stichprobe	willkürlich, zufällig, quasi zufällig	Im Ermessen des Prüfers unter Berücksichtigung des Stichprobenrisikos

2.3.2 Besonderheiten von Grundgesamtheiten in der Jahresabschlussprüfung

Die klassische Stichprobentheorie beruht auf der Annahme der Normalverteilung der zugrundeliegenden Daten, wobei lange Zeit von der Übertragbarkeit auf Problemstellungen in der Abschlussprüfung ausgegangen wurde[19]. Verteilungsannahmen dienen dabei der Beschreibung von empirisch beobachtbaren Häufigkeitsverteilungen, zu denen auch Fehleranteile bzw. Fehlerumfänge im Prüfungswesen gehören. Durch Schätzung von Verteilungsparametern auf Basis einer Stichprobe wird eine Verteilungsfunktion modelliert und so die Verteilung des wahren Fehleranteils bzw. der wahren Fehlerhöhe geschätzt. Da unter Annahme der Normalverteilung eine Vielzahl von Verfahren für Stichprobenprüfungen in der Abschlussprüfung adaptiert werden konnte, war eine solche Annahme zu favorisieren. Jedoch wurden zeitnah Beweise dahingehend geliefert, dass die Verteilung

[17] Vgl. dazu auch IDW (Hrsg.), a.a.O. 2012, R, Tz. 120, S. 2433.
[18] Eigene Darstellung in Anlehnung an *Guy/Carmichael/Whittington*, a.a.O. 2002 S. 221.
[19] Vgl. zusammenfassend *Fienberg/Neter/Leitch*, JASA 358/1977, S. 295.

von Fehlern und die Varianz von Fehlergrößen in realen Prüffeldern i.d.R. nicht der Normalverteilung entsprechen; vielmehr sind Grundgesamtheiten in der Abschlussprüfung durch eine hohe Schiefe und geringe Fehlerraten gekennzeichnet[20]. Diese Erkenntnis war ausschlaggebend dafür, dass alternative Verfahren, die relativ unabhängig von der zugrundeliegenden Verteilung arbeiten, entwickelt worden sind, darunter insbesondere das MUS.

2.3.3 Auswahlprüfungen im Lichte der Berufsgrundsätze

Auswahlprüfungen bergen die Gefahr, dass Grundsätze ordnungsmäßiger Abschlussprüfung unbewusst nicht eingehalten werden. Insbesondere die Gewissenhaftigkeit bei der Auftragsplanung sowie die Verantwortlichkeit der eigenen Urteilsbildung und eigenen Urteilsfindung spielen für Auswahlverfahren eine besondere Rolle[21]. Dabei geht es um die Sicherstellung der Beachtung gesetzlicher, beruflicher und fachlicher Bestimmungen, so dass eine Einschränkung der anwendbaren Verfahren bereits per se vorliegt. Der Prüfer hat eigenverantwortlich aus dem Fundus verfügbarer Verfahren auszuwählen und diese im Hinblick auf Zweckadäquanz und Mitteleinsatz zu würdigen[22].

2.3.4 Einsatzmöglichkeiten von Auswahlprüfungen

Grundsätzlich lassen sich solche Prüfungsziele unterscheiden, die eine Aussage hinsichtlich des Fehleranteils (homograder Fall) anstreben und solche, bei denen die Fehlerhöhe (heterograder Fall) im Fokus steht. Während Funktionsprüfungen die Fehlerhäufigkeit bzw. den Fehleranteil betrachten, verfolgen darauf aufbauende Einzelfallprüfungen auch die Fehlerhöhe, die direkt in Beziehung zur quantitativen Wesentlichkeit gesetzt werden kann.

Der Output des betrieblichen Prozesses des Rechnungswesens sind i.d.R. Buchwerte. Eine statistische Ergebnisprüfung stellt somit regelmäßig diejenige von Fehlerwerten dar[23]. Nach herrschender Meinung ist der Einsatz von Auswahlprüfungen in diversen Prüffeldern obligatorisch[24], darunter bei der

[20] Vgl. z.B. *Durney/Elder/Glover*, Auditing-J Pract Th 2/2013, S. 79-110; *Hömberg*, BFuP 3/1997, S. 250; *Johnson/Leitch/Neter*, TAR 2/1981, S. 283.
[21] Vgl. §§ 43, 53, 55a WPO; §§ 11, 12 BS.
[22] Vgl. *Graumann*, Wirtschaftliches Prüfungswesen, Herne 2015, S. 47 f.
[23] Vgl. *Hömberg*, a.a.O. 1997, S. 249.
[24] Vgl. *Christensen/Elder/Glover*, Account Horiz 1/2015, S. 61-81.

- Prüfung sämtlicher Bereiche des internen Kontrollsystems (Funktionsprüfungen),
- Prüfung von Forderungen und Verbindlichkeiten (z.B. Saldenbestätigungsaktion),
- Prüfung der Vorräte (Inventurbeobachtung, Prüfung des Mengengerüsts, Preistests),
- Prüfung von Erlös- und Aufwandsposten (Realisierung, Periodenabgrenzung).

Im homograden Fall wird der Prüfer ein Prüffeld dabei als nicht ordnungsmäßig zurückweisen, wenn die geschätzte Anzahl der Fehler mit einer hohen Wahrscheinlichkeit einen vorzugebenden Wert überschreitet. Der in der Stichprobe identifizierte Fehleranteil entspricht dabei demjenigen, der auch für die Grundgesamtheit erwartet wird. Im heterograden Fall wird der Prüfer ein Prüffeld als nicht ordnungsmäßig bezeichnen, wenn die Über- oder Unterbewertung mit einer hohen Wahrscheinlichkeit einen vorzugebenden Wert überschreitet. Dazu rechnet der Prüfer die durch ein Stichprobenergebnis geschätzte Fehlerhöhe auf die Grundgesamtheit hoch. Dies geschieht durch einen Soll-Ist-Vergleich der vom Unternehmen angesetzten Buchwerte mit dem vom Prüfer festgestellten und nach den Grundsätzen ordnungsmäßiger Bilanzierung anzusetzenden Sollwert. Bei korrekter Buchung stimmen die beiden Werte überein, ansonsten liegt eine Differenz vor. Aufgrund der Tatsache, dass im heterograden Fall auch der Umfang individueller Fehler betrachtet wird, stellen sich die Wahl des Erhebungsverfahrens sowie die Art der Hochrechnung ermittelter Soll-Ist-Differenzen ungleich komplexer dar.

2.3.5 Nutzung von Vorinformationen

Der risikoorientierte Prüfungsansatz verlangt vom Prüfer die zielgerichtete Beschaffung von Vorinformationen. So werden zunächst sämtliche das Risiko beeinflussende Informationen ausgewertet, ehe eine Ergebnisprüfung durchgeführt wird[25]. IDW EPS 300 n.F. sieht vor, diese Vorinformationen zur Determinierung des notwendigen Umfangs von Prüfungsnachweisen zu nutzen[26]. Da Stichproben erst deutlich nach Prüfungsbeginn genutzt werden, ist die Ungewissheit zum Einsatzzeitpunkt bereits reduziert. Es können so Annahmehypothesen formuliert werden, deren Untersuchung über den Verlauf der Prüfung zu Prüfregeln führt. Beispiele:

[25] Vgl. *Wolz*, Wesentlichkeit im Rahmen der Jahresabschlussprüfung, Düsseldorf 2003, S. 33.
[26] Vgl. IDW EPS 300 n.F., Anlagen 2, 3.

a) *Wenn* das interne Kontrollsystem zuverlässig funktioniert, *dann* kann es eigentlich zu keinen schwerwiegenden Fehlern kommen, *so dass* die Menge der Einzelfallprüfungshandlungen geringer bemessen werden kann.
b) *Wenn* Mitarbeiter A aus der Debitorenbuchhaltung aufgrund von Krankheit im Juli vom weniger erfahrenen Mitarbeiter B vertreten wurde, *dann* erhöht sich für diesen Zeitraum die Wahrscheinlichkeit wesentlicher Fehler, *so dass* die Anzahl der Einzelfallprüfungshandlungen für diesen Zeitraum zu erhöhen ist.

Ein guter Überblick über die Prüfungssituation und eine angemessene Einschätzung des Prüfers führen dazu, dass sich die Auswahl tendenziell von statistischen Methoden entfernen darf[27]. Da die Anzahl benötigter Vorinformationen bei einer Zufallsauswahl generell geringer ist, bietet sich letztere vor allem für Systemtests sowie umfangreiche monetäre Prüffelder an[28]. Eine sinnvolle Kombination ist in folgender Art und Weise denkbar und in der Regel hinreichend effizient:

[27] Vgl. *Baetge*, Auswahlprüfungen auf der Basis der Systemprüfung, in: o.A. (Hrsg.), Wirtschaft und Wissenschaft im Wandel. Festschrift für Carl Zimmerer zum 60. Geburtstag, Frankfurt 1986, S. 53.
[28] Vgl. *Swearingen/Hansen*, JABR 4/1990, S. 53, die zeigen, dass hinsichtlich der Effizienz bei der Fehlersuche nichtstatistische Verfahren durchschnittlich effektiver sind als statistische Verfahren.

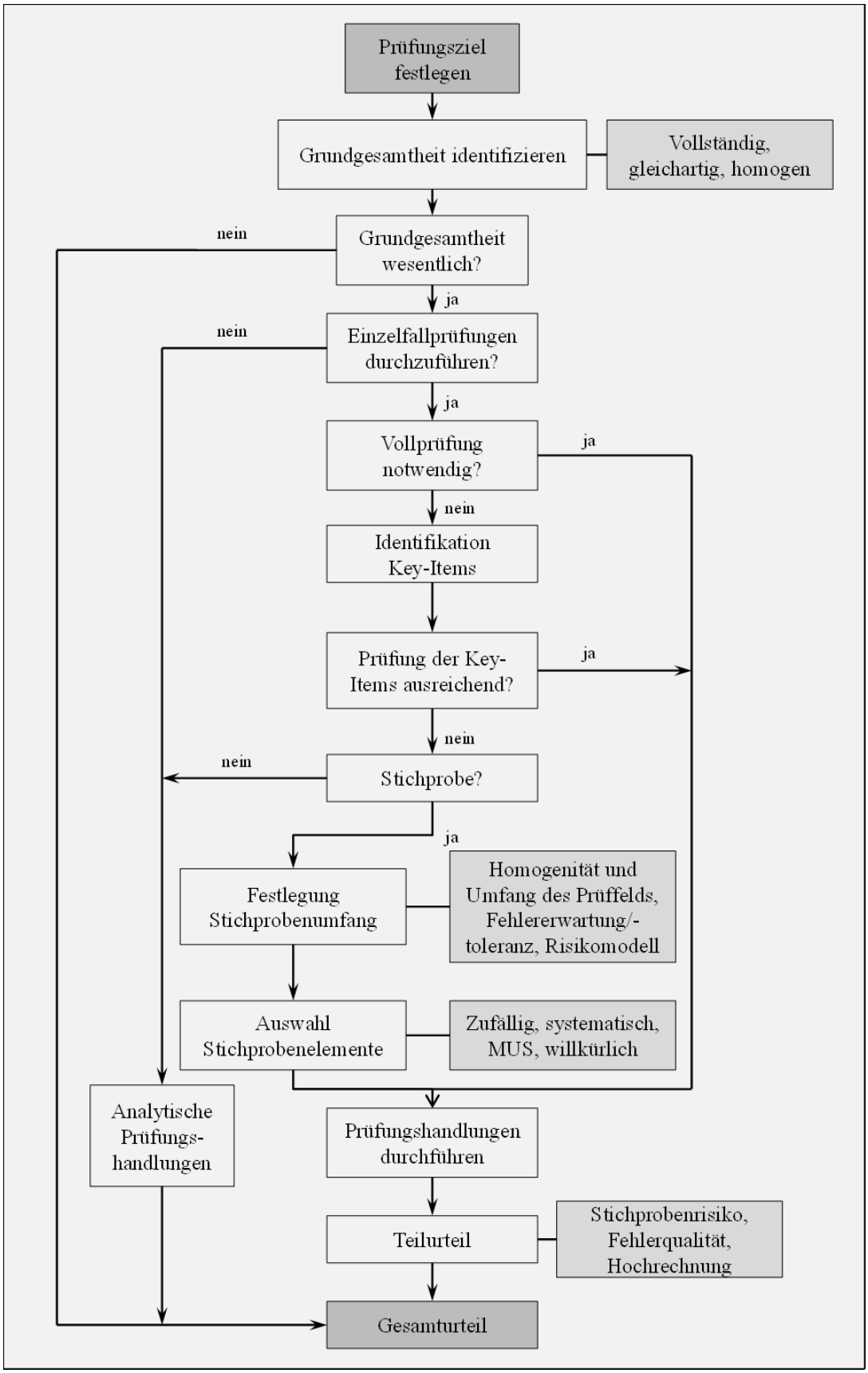


Abbildung 2.2: Beispielhafter Ablauf einer Stichprobenprüfung[29]

[29] Eigene Darstellung in Anlehnung an *Barnett/Read*, J Accountancy 1/1986, S. 79.

Nichtstatistische Stichproben sind in einigen Fällen notwendig und häufig die einzige Möglichkeit für effektive und effiziente Prüfungshandlungen. Beispiele:

- Die Auswahl zu zählender bzw. zu wiegender Elemente bei der Inventurbeobachtung hängt von der Beschaffenheit des Lagerortes, der Einhaltung von Inventurrichtlinien sowie der Genauigkeit der Testzählungen ab. Eine Vorabbestimmung des Stichprobenumfangs ist dahingehend nicht hinreichend und muss iterativ festgelegt werden.
- In sensiblen Prüffeldern wie Rechts- und Beratungskosten wird der Prüfer häufig bewusst auswählen wollen, welche Elemente einer Untersuchung zu unterziehen sind. Da hierbei Erfahrungen und Erkenntnisse eine wichtige Rolle spielen, ist die Anwendung einer Zufallsauswahl nicht zielführend.
- Bei der gemeinsamen Betrachtung mehrerer Konten aufgrund von Wesentlichkeitsüberlegungen ist eine statistische Erhebung häufig nicht möglich. So können in den Bereichen Rechtsberatung, externe Beratung, externe Buchführung und Steuerberatung einzelne Buchwerte immateriell sein, die zusammen gesehen jedoch über der Wesentlichkeit liegen.

2.4 Methoden der Auswahlprüfung

Regelmäßig wird von der Fehlerfreiheit bzw. einer geringen Fehlerquote der Grundgesamtheiten in Jahresabschlüssen ausgegangen. So weisen Verteilungen von Buchwerten und Abweichungen teils extreme Eigenschaften auf, die mit Hilfe üblicher statistischer Verfahren nicht angemessen berücksichtigt werden können. Gemäß den wenig leitenden Anforderungen aus Gesetz und Literatur hat der Prüfer sein Urteil grundsätzlich auf Basis von Prüfungsmethoden abzugeben, die dem Prüfungszweck dienlich sind[30], wobei häufig auf bestehende und vermeintlich effektive Prüfungsmethoden zurückgegriffen wird[31]. Praxisrelevante (Auswahl-)Verfahren neben der bewussten Auswahl sind im Wesentlichen die in IDW EPS 310, Anlage 4 genannten:

- Zufallsauswahl (echte oder unechte, z.B. mittels Zufallszahlengenerator)
- Systematische Auswahl (mit Zufallsstart)

[30] Vgl. zur Präferierung nichtstatistischer Methoden in der internationalen Prüfungspraxis z.B. *Hitzig*, The CPA Journal 7/1995, S. 54-57, *Elder/Allen*, Auditing-J Pract Th 1998, S. 71-87; *Durney/Elder/Glover,* a.a.O.; *Elder/Akresh/Glover/Liljegren*, Auditing-J Pract Th 2013, Suppl. 1, S. 99-129.

[31] Vgl. zu dieser Problematik bei analytischen Prüfungshandlungen auch *Bierkämper/Toll*, WP Praxis 10/2015, S. 254.

- MUS (wertproportionale Auswahl)
- Zufallsimitierende Auswahl (*haphazard sampling* bzw. willkürliche Auswahl)
- Blockauswahl (zusammenhängende Elemente, z.B. alle Umsatzerlöse im Dezember)

Diese Aufzählung ist im Sinne zur Verfügung stehender Verfahren nicht abschließend, stellt jedoch die gebräuchlichen und in den meisten Softwaretools zur Verfügung stehenden Verfahren dar.

2.4.1 Bewusste Auswahl

Da die Prüfungsstandards lediglich die Wahl eines für den jeweiligen Prüfungszweck effektiven Auswahlverfahrens verlangen, bietet es sich im Rahmen des risikoorientierten Prüfungsansatzes[32] an, solche Elemente auszuwählen, bei denen der Prüfer bspw. aufgrund von Vorjahresergebnissen oder während der Prüfung gewonnener Erkenntnisse von einer erhöhten Fehleranfälligkeit ausgehen muss. Er wird sich unter gegebenen Umständen nicht davon abbringen lassen, gewisse Elemente in jedem Fall einer Prüfung zu unterziehen. Dabei ist zu unterscheiden, ob diese Elemente isoliert betrachtet werden, während die restlichen Elemente z.B. auf Basis einer Zufallsauswahl gesondert gewürdigt werden, oder aber sämtliche Elemente im Prüffeld bewusst ausgewählt werden. Der erste Fall würde einer Schichtung der Grundgesamtheit entsprechen, sodass weiterhin von einer Stichprobenprüfung die Rede sein kann; im zweiten Fall handelt es sich im Sinne des IDW EPS 300 n.F. nicht um eine Stichprobe. Dem Prüfer muss in einem solchen Fall bewusst sein, dass er sich dem Zustand aussetzt, keine Aussagen zu den Eigenschaften der verbleibenden Elemente im Prüffeld treffen zu können[33].

Von Stichproben i.e.S. abzugrenzen sind zukünftig solche Auswahlprüfungen, bei denen bestimmte Elemente mit einer erhöhten Wahrscheinlichkeit durch Einflussnahme des Prüfers selektiert werden (bewusste Auswahl). Dies können sein[34]:

- Elemente mit hohem Wert oder Schlüsselelemente,
- alle Elemente, die einen bestimmten Betrag überschreiten,
- Elemente zum Erlangen von Informationen.

[32] Vgl. § 317 Abs. 1 Satz 3 HGB; IDW PS 261, Tz. 5 ff.
[33] Vgl. *Burgstahler/Glover/Jiambalvo*, Auditing-J Pract Th 1/2000, S. 79-99; *Elder/Allen*, a.a.O. 1998, S. 81.
[34] IDW EPS 300 n.F., Tz. A50.; ISA 530.A52; vgl. kritisch auch *Mochty*, Nichtstatistische Stichproben, in: *Kirsch* (Hrsg.), Rechnungslegung und Wirtschaftsprüfung. Festschrift zum 70. Geburtstag von Jörg Baetge, Düsseldorf 2007, S. 1060.

Allen o.g. Methoden ist gemein, dass die verbleibenden Elemente der Grundgesamtheit unterrepräsentiert sein werden und demnach keine Wissensbasis bestehen kann, die eine Aussage über alle Elemente der Grundgesamtheit ermöglicht[35]. Eine quantitative Hochrechnung verbietet sich entsprechend. Es bleibt darüber hinaus zu beachten, dass auch Posten mit geringen Werten kumuliert wesentliche Fehler enthalten können.

Im Rahmen der Dokumentation der Prüfungshandlungen einer bewussten Auswahl darf in den Arbeitspapieren künftig nicht mehr von der Ziehung einer Stichprobe die Rede sein. Es muss klar ersichtlich sein, dass die Auswahl von Untersuchungsgegenständen nicht auf Basis unabhängiger Ziehungsmethoden geschehen ist und kein Anspruch erhoben wird, auf Basis dieser Prüfungshandlung Aussagen über das gesamte Prüffeld machen zu können. Insbesondere auch im Hinblick auf die interne und externe Qualitätssicherung und damit auf die Exkulpation ggü. Dritten sind diese Auswirkungen auf die Prüfungsdokumentation essenziell. Nichtsdestotrotz bietet sich in vielen Fällen die bewusste Auswahlmethodik an, da sie in hohem Maße die Nutzung von Vorinformationen und Prüferwissen erlaubt und im Sinne der Prüfungsstandards eine adäquate Herangehensweise zur Erlangung von Prüfungsnachweisen darstellt.

2.4.2 Statistische Stichprobenverfahren

2.4.2.1 Überblick und Anwendungsvoraussetzungen statistischer Stichprobenverfahren

Zu den statistischen Stichprobenverfahren gehören gemäß IDW EPS 310, Tz. A11 (nur) solche, die jede Art von systematischer Verzerrung vermeiden und so einen Repräsentationsschluss zulassen. Von praktischer Relevanz sind dabei vor allem die einfache Zufallsauswahl, die geschichtete Zufallsauswahl sowie das MUS.

Die Zufallsauswahl ist bereits zum jetzigen Zeitpunkt für alle Prüfungen legitim[36]. Bei den entsprechenden Verfahren werden die Stichprobenelemente so ausgewählt, dass der Prüfer keinerlei Einfluss auf die Auswahlwahrscheinlichkeiten und die gezogenen Elemente hat; sie werden damit zu statistisch verwertbaren Zufallsvariablen. Die Auswahl erfolgt i.d.R. mittels EDV-Unterstützung entweder per echter oder unechter

[35] Vgl. IDW EPS 300 n.F., Tz. A51; ISA 500.A54.
[36] Vgl. IDW St/HFA 1/1988, Abschn. D.I.

Zufallsauswahl[37]. Die bedeutenden Vorteile einer statistischen Auswertung von Stichproben sind die Quantifizierung des Stichprobenrisikos bzw. der erreichten Prüfungssicherheit sowie die Möglichkeit der Berechnung des Grenznutzens zusätzlich gezogener Stichprobenelemente. Es kann in Abhängigkeit des Stichprobenumfangs, der Anzahl der Stichprobenelemente und ggf. des Umfangs gefundener Fehler sowie der prüffeldspezifischen Wesentlichkeit angegeben werden, mit welcher Wahrscheinlichkeit ein Sollwert in einem bestimmten Intervall um den wahren Istwert liegt[38].

Wenngleich Softwarelösungen wie IDEA oder ACL die Anwendung statistischer Methoden erlauben und lediglich die Eingabe einiger weniger Parameter verlangen, wird für die Praxis ausdrücklich davor gewarnt, statistische Methoden ohne ein grundlegendes Verständnis der zugrundeliegenden Annahmen anzuwenden. Aufgrund der stets subjektiven Wahl von Parametern wie Fehlererwartung, Fehlertoleranz sowie dem gewählten Konfidenzniveau (äquivalent zur Prüfungssicherheit) besteht ein nicht zu unterschätzendes Risiko, anhand vermeintlich objektiver und unabhängig erlangter Ergebnisse zu einem nicht zutreffenden Urteil zu gelangen[39]. Nicht zuletzt aufgrund der wenig universellen Einsetzbarkeit sind z.B. die gebundenen Schätzverfahren (Differenzen-, Verhältnis- und lineare Regressionsschätzung) aus dem Instrumentarium nahezu vollständig verschwunden[40].

Allen statistischen Stichprobenverfahren ist gemeinsam, dass für eine effektive und effiziente Anwendung einige Anforderungen an die Grundgesamtheit erfüllt sein sollten. So muss diese einen gewissen Umfang besitzen (der AICPA Audit Guide „Audit Sampling" spricht in diesem Zusammenhang durchweg von *large populations*; folglich wird eine Anwendung für kleine Prüffelder bereits formal scheitern). In der Literatur existieren verschiedene Ansätze; zur Nutzung der statistischen Inferenz sollte jedoch ein Umfang von 500 oder mehr Elementen vorliegen[41].

Daneben muss die Grundgesamtheit hinreichend homogen sein, d.h. die Elemente müssen formal und sachlich vergleichbar sein, um statistisch evaluiert werden zu können. Formale Vergleichbarkeit bedeutet, dass alle Elemente z.B. Belege oder Buchungen sind. Dies ist i.d.R. problemlos beobachtbar. Zentraler ist die sachliche Vergleichbarkeit: Das

[37] Vgl. für die Darstellung verfügbarer Ziehungsverfahren z.B. *von Wysocki*, a.a.O. 2002, S. 199 f.; *Weinand/Wolz*, a.a.O. 2012, S. 69 f.
[38] Vgl. z.B. *Hömberg*, a.a.O. 1997, S. 250.
[39] Vgl. IDW (Hrsg.), a.a.O. 2012, R, Tz. 130, S. 2434-2435.
[40] Vgl. hierzu *Giezek*, a.a.O. 2014, S. 567.
[41] Vgl. z.B. *Hitzig*, a.a.O. 2004, S. 31.

zu untersuchende statistische Merkmal muss über die Grundgesamtheit gleichartig sein und in verschiedenen Ausprägungen vorliegen, d.h. eine gewisse Varianz besitzen (da es sich dabei i.d.R. um monetäre Werte handelt, kann diese Annahme grundsätzlich als gegeben angesehen werden).

Auch wenn statistische Methoden das Auswahlverfahren objektivieren, bleibt die Bestimmung von Prüfregeln sowie die Interpretation erlangter Ergebnisse, inklusive der daraus resultierenden Effekte auf weitere Prüfungshandlungen, in den Händen des Prüfers und ist durchweg kontextabhängig. In der Literatur wird teilweise vorgeschlagen, die Verfahrenswahl auch von den Fähigkeiten des Prüfers, der Ressourcenausstattung der Prüfungsgesellschaft und der vorhandenen Software abhängig zu machen. Zuweilen wird sogar favorisiert, statistische Methoden zu ignorieren, wenn kein Prüfer mit statistischem Hintergrundwissen verfügbar ist[42]. Diese Aussagen sind in höchstem Maße kritisch zu sehen, da die Wahl einer Methode stets dem Prüfungszweck dienlich sein muss und mangelnde Sachkenntnis kein Argument für die Verwendung ineffektiver Verfahren sein darf.

2.4.2.2 Monetary Unit Sampling

Auswahlverfahren mit Verteilungsannahmen sind für das Prüfungswesen oft unbrauchbar oder schlicht unwirtschaftlich. Die klassische statistische Methodenlehre hält jedoch keine für die Abschlussprüfung praktikablen Alternativen bereit. Aus diesem Grund wurden Verfahren entwickelt, die den Eigenschaften von Grundgesamtheiten in der Abschlussprüfung besser gerecht werden. Das mit Abstand bedeutendste Verfahren (genutzt für vier von fünf Stichproben[43]) ist dabei das MUS[44], welches die homograde und heterograde Fragestellung kombiniert[45]. Das Verfahren ist in nahezu allen Softwarepaketen zur Stichprobenerhebung enthalten. Seine flächendeckende Nutzung begründet sich weniger in der praktischen Effizienz als vielmehr in der relativ universellen Einsetzbarkeit.

IDW EPS 310, Tz. A8 erlaubt es dem Prüfer, eine solche wertproportionale Stichprobenauswahl vorzunehmen (*probability proportional to size sampling*[46]). Diese Methodik entspricht dem o.g. MUS, welches sich durch eine maximale Schichtung der

[42] Vgl. *Colbert*, JABR 2/1991, S. 120.
[43] Vgl. *Newiak*, Prüfungsurteile mit Dollar Unit Sampling, Potsdam 2009, S. 1.
[44] Vgl. *Leslie/Teitlebaum/Anderson*, Dollar Unit Sampling, Toronto 1979.
[45] Vgl. *Giezek*, Monetary Unit Sampling, Wiesbaden 2011, S. 4.
[46] Vgl. AICPA (Hrsg.), a.a.O. 2014, 4.17; ISA 530, Anlage 1.

Grundgesamtheit auszeichnet[47]. Diese wird erreicht, indem anstatt der Einzelposten des Prüffelds (z.B. Konten oder Belege) die *Monetary Units* i.S. einzelner Geldeinheiten im Prüffeld betrachtet werden; ein Prüffeld mit einem Gesamtbetrag von € 102.350 besteht dementsprechend aus ebenfalls 102.350 Elementen, denen gleiche Ziehungswahrscheinlichkeiten zugeordnet werden. Es werden dann die mit der ausgewählten Geldeinheit korrespondierenden Transaktionen oder Belege als konkrete Untersuchungsobjekte geprüft. Wird davon ausgegangen, dass die zu erwartenden Fehler (i.d.R. Überbewertungen) mit den Buchwerten korrelieren, ergibt sich so eine implizite Berücksichtigung der zugrunde liegenden Verteilung und eine effektive – wenngleich im Gegensatz zur gebundenen Hochrechnung weniger effiziente[48] – Auswahlmethode. Das Ziehen der Stichprobenelemente geschieht ohne Zurücklegen. Dennoch kann ein Untersuchungsgegenstand mit einem Buchwert größer € 1 mehrmals gezogen werden. Dadurch, dass mehrere Geldeinheiten auf ein und denselben Posten entfallen, wird die tatsächlich zu prüfende Menge der Elemente damit regelmäßig geringer als der seitens der Prüfungssoftware a priori berechnete Stichprobenumfang ausfallen.

IDW EPS 310 weist in Anlage 1, Nr. 5 darauf hin, dass mit Hilfe des MUS eine Fokussierung auf höherwertige Elemente in der Grundgesamtheit stattfindet. Durch diese einseitige Gewichtung eignet es sich damit insbesondere für die Prüfung von Prüffeldern, bei denen ein Verdacht auf Überbewertungen vorliegt[49]. Unterbewertete Elemente erhalten folglich eine geringere Wahrscheinlichkeit, in die Stichprobe zu gelangen[50], Nullsalden finden keine Berücksichtigung und sollten gesondert betrachtet werden.

2.4.3 Nichtstatistische Stichprobenverfahren

Im Gegensatz zum in der Literatur bisher häufig vertretenen Verständnis, bei nichtstatistischen Verfahren handele es sich um solche, denen keine Zufallsauswahl – sondern eine bewusste Auswahl – zugrunde liegt, stellen die neuen Prüfungsstandards konvergent mit den entsprechenden ISAs die Abgrenzung klar. IDW EPS 300 n.F., Tz. A51 erläutert in Entsprechung des ISA 500.A54 zur nichtstatistischen Stichprobe, dass eine selektive Untersuchung (entspricht der bewussten Auswahl) von Elementen häufig ein

[47] Vgl. *Weinand/Wolz*, WP Praxis 1/2013, S. 13-17 für eine Darstellung der mathematischen Grundlagen der buchwertproportionalen Auswahl sowie ein Anwendungsbeispiel zur Prüfung von Forderungsposten.
[48] Vgl. *Jung/Kellerer*, RWZ 8/1995, S. 248, Newiak; a.a.O. 2009, S. 24, 28, 50 f.
[49] Vgl. *Christensen/Elder/Glover*, a.a.O. 2015, S. 75 zur Problematik, dass das MUS teilweise willkürlich auch zur Prüfung von Verbindlichkeitsposten angewendet wird.
[50] Vgl. demgegenüber *Wolz*, BFuP 1/2004, S. 60 f.

wirtschaftliches Mittel zur Erlangung von Prüfungsnachweisen ist, dabei jedoch keine Stichprobenprüfung darstellt, da sich eine Hochrechnung der Ergebnisse auf die Grundgesamtheit aufgrund statistischer Restriktionen verbietet. Um ein nichtstatistisches Stichprobenverfahren handelt es sich demnach immer dann, wenn mindestens eine der Voraussetzungen des IDW EPS 310, Tz. 7g) verletzt ist. Hierunter wird insbesondere der in praxi häufig anzutreffende Fall subsumiert, dass die Auswahl der Stichprobenelemente zufällig, die Ergebnisinterpretation jedoch nach prüferischem Ermessen geschieht.

IDW EPS 310, Tz. A9 erlaubt dem Prüfer, anders als IDW St/HFA 1/1988, dabei explizit die Nutzung der zufallsimitierenden Auswahl (*haphazard sampling*), bei der die Zufallsauswahl durch den Prüfer nachzuahmen versucht wird. Dabei wird für die Praxis eindeutig von dieser Methode abgeraten, da von einer unverzerrten Zufallsziehung durch den Prüfer nicht ausgegangen werden kann, da sich dieser stets von strukturellen und visuellen Reizen in der Grundgesamtheit beeinflussen lässt[51].

Insgesamt wird damit in Zukunft vermehrt auch auf die Verwendung mathematisch-statistischer Termini in den Arbeitspapieren zu achten sein, um in Zweifelsfällen keine Prüfungsurteile zu formulieren, die anhand der verwendeten Methode nicht erreichbar gewesen wären.

2.5 Fazit und Ausblick

Das IDW folgt mit der Veröffentlichung von IDW EPS 300 n.F. und IDW EPS 310 konsequent den Vorgaben des IFAC zur Erlangung von Prüfungsnachweisen sowie der Anwendung von Stichprobenverfahren im Rahmen der Jahresabschlussprüfung. Auch nach Verabschiedung der neuen Prüfungsstandards wird dabei weder der Einsatz der bewussten Auswahl zu prüfender Elemente eingeschränkt, noch wird der Einsatz statistischer Stichprobenverfahren verbindlich. Es ändert sich jedoch die Betrachtungsweise von Auswahlprüfungen seitens des Standardsetters. Der Weg einer methodenorientierten Betrachtungsweise wird zugunsten eines praxisorientierten Ansatzes verlassen, indem der Darstellung wesentlicher Voraussetzungen zur Stichprobenprüfung und der Fokussierung auf wesentliche Problemfelder Vorzug gegeben wird. Eine Konkretisierung der Verfahrensanwendung bleiben die Standards schuldig. Damit verbleibt die Frage nach der

[51] Vgl. empirisch *Hall/Higson/Pierce/Price/Skousen*, BRIA 2/2012, S. 101-132; IDW (Hrsg.), a.a.O. 2012, R, Tz. 122, S. 2433; *Hill*, Psychol Rep 3/1988, S. 967-971; *Mochty*, Stichprobentechnik für Statistik-averse Wirtschaftsprüfer, in: *Schröder/Clausen/Behr* (Hrsg.), Essener Beiträge zur empirischen Wirtschaftsforschung, Festschrift für Prof. Dr. Walter Assenmacher, Wiesbaden 2012, S. 76.

praktischen Umsetzung der normativen Vorgaben. Insbesondere die Vermeidung wesentlicher Fallstricke bei der praktischen Durchführung von Auswahlprüfungen wird i.R.d. zukünftigen Normen zwar angedeutet, die tatsächliche Ausgestaltung bleibt jedoch der professionellen – und im Zweifel höchst subjektiven – Einschätzung des Prüfers überlassen. Die Entwicklung einer allgemeinen Strategie zur verfahrensunabhängigen Berücksichtigung scheint aus diesem Grunde nicht nur wünschenswert, sondern notwendig und soll in einem weiteren Artikel näher betrachtet werden.

3 Die praktische Durchführung von Auswahlprüfungen – Strategien zur Vermeidung wesentlicher Fallstricke bei der Erhebung und Auswertung von Stichproben

3.1 Publikationsdetails

Zusammenfassung: Entscheidet sich der Prüfer bei der Prüfungsplanung zur Verwendung von Auswahlprüfungen, hat er bei der schlussendlichen Durchführung einerseits die den verfügbaren Auswahlverfahren immanenten Anwendungsvoraussetzungen zu berücksichtigen und andererseits die Beschaffenheit der zu untersuchenden Daten zu würdigen. Nachdem mit der Veröffentlichung von IDW EPS 310 zur Umsetzung der Anforderungen des IFAC bzgl. der Durchführung von repräsentativen Stichprobenprüfungen i.S.d. ISA 530 zwar eine Sensibilisierung für spezifische Problemfelder stattgefunden hat, verbleibt die Frage nach der praktischen Vermeidung der dort genannten Fallstricke. Ziel dieses Beitrags ist es daher die wesentlichen Problembereiche aufzuzeigen und zu diskutieren und dem geneigten Prüfer so eine geeignete Wissensbasis zur Entwicklung einer verfahrensunabhängigen Vermeidungsstrategie zur Verfügung zu stellen.

Koautoren: Christoph Oldewurtel.

Stichwörter: Auswahlprüfung, Nichtstatistische Stichproben, Stichprobenrisiko, Fehlerisolierung, Stichprobenumfang.

Publikationsstatus: Veröffentlicht in: *WP Praxis* 5 (8): 199–204.

3.2 Auswahlprüfungen im Sinne geltender und zukünftiger Prüfungsnormen

Durch die Verlagerung in IDW EPS 300 n.F. wird der bewussten Auswahl der Status als Stichprobenverfahren i.S.d. IDW St/HFA 1/1988 aberkannt[52]. Letztere Stellungnahme des HFA wird dahingehend durch den im Mai 2015 veröffentlichten IDW EPS 310 ersetzt, im Rahmen dessen die Darstellung von Grundlagen zur Anwendung repräsentativer statistischer und nichtstatistischer Stichprobenverfahren in Entsprechung mit ISA 530 stattfindet. In der praxisorientierten Literatur besteht dabei weithin keine Einigkeit zur Vorteilhaftigkeit statistischer Stichprobenverfahren[53]. Mithin scheint es jedenfalls schwierig, die Vorgaben von ISA 530 und IDW EPS 310 hinsichtlich der Notwendigkeit statistisch motivierter Verfahren zu erkennen, was in der Literatur zu teils stark differierenden Sichtweisen führt[54].

Durch die wenig verfahrensbezogene Ausrichtung sanktionieren die Prüfungsstandards weniger die reine Methodenwahl als vielmehr eine mangelnde Berücksichtigung statistischer Gesetzmäßigkeiten und prüffeldspezifischer Eigenschaften. Eine Förderung der Anwendung statistischer Methoden ist in IDW EPS 310 damit nicht zu sehen. Im Vergleich zur IDW St/HFA 1/1988 ist sogar von einer methodisch weniger umfangreichen Darstellung zu sprechen. Auswahlprüfungen bestehen dabei aus mehreren Schritten, deren konkrete Ausgestaltung von der Beschaffenheit des Prüffelds, dem Umfang belastbarer Vorinformationen sowie den angestrebten Aussagen der Prüfungshandlung abhängt. Den Prüfungsstandards des IFAC und des IDW ist dabei gemein, dass zwar ein zu erreichendes Ergebnis definiert wird, eine hinreichend genaue Beschreibung zur Herleitung dieses Ergebnisses jedoch ausbleibt. Stattdessen wird die Aufmerksamkeit des Prüfers auf spezifische das Stichprobenrisiko im Besonderen determinierende Faktoren gelenkt,

[52] Vgl. *Baumeister/Oldewurtel*, WP Praxis 07/2016, S. 169.

[53] Vgl. dazu kritisch *Mochty*, Stichprobentechnik für Statistik-averse Wirtschaftsprüfer, in: *Schröder/Clausen/Behr* (Hrsg.), Essener Beiträge zur empirischen Wirtschaftsforschung, Festschrift für Prof. Dr. Walter Assenmacher, Wiesbaden 2012, S. 76.

[54] Vgl. dazu kontrovers z.B. *Göb/Karrer*, WPg 11/2010, S. 600, 602: „Die Vorgaben von ISA 500 und ISA 530 machen den Einsatz der statistischen Stichprobenprüfung in der Wirtschaftsprüfung unabdingbar."; „Bei der Übernahme der ISA bzw. deren Transformation in nationale Prüfungsstandards kann die weit verbreitete bisherige Berufspraxis nicht fortgeführt werden. Die bewusste Auswahl ist keine Alternative zur Stichprobenprüfung. […] Zur Unterstützung der unausweichlich gewordenen statistischen Stichprobenprüfung müssen […] neue Hilfsmittel in die Berufspraxis Einzug halten."; demgegenüber *Heese/Braatsch*, WPg 17/2013, S. 848: „ISA 530 beschreibt mit repräsentativen Stichprobenverfahren ein in der Prüfungspraxis begrenztes Anwendungsfeld."; „Eine mathematisch-exakte Definition des Stichprobenrisikos erscheint hier in den meisten Praxisfällen nicht erforderlich […].".

wobei die Standards eine explizite Taktik zur Vermeidung und zum Umgang mit diesen Problemen schuldig bleiben.

3.3 Wesentliche Fallstricke bei der Durchführung von Auswahlprüfungen

Aufbauend auf der Annahme, dass sich der Prüfer zur Untersuchung des Jahresabschlusses regelmäßig für den Einsatz von Stichprobenprüfungen bzw. Auswahlprüfungen im Allgemeinen entscheidet, muss er verfahrensunabhängig stets die Besonderheiten von Grundgesamtheiten im Jahresabschluss berücksichtigen, da diese häufig nicht die Eigenschaften üblicher normalverteilter Daten widerspiegeln und wesentliche Anwendungsvoraussetzungen zur statistischen Auswertung deshalb oftmals nicht gegeben sind[55]. Der Prüfer hat gem. § 11 Abs. 2 BS alle Tätigkeiten zu unterlassen, bei denen er seine eigene Urteilsbildungsfähigkeit nicht sicherstellen kann. Hierzu gehört demnach auch die Anwendung eines den an die geplante Prüfungshandlung gestellten Anforderungen nicht entsprechenden oder die Nutzung eines ihm nicht vertrauten Erhebungsverfahrens. Aufgrund der durchweg softwarebasierten Anwendung vermag ein Auswahlverfahren damit zwar anwendbar sein, ohne grundlegende statistische Kenntnisse vermag es jedoch zur „*Black Box*“ zu werden. Wenngleich mit IDW EPS 310 in Ansätzen eine Sensibilisierung für ebensolche Problemfelder stattfindet, erfolgt im Folgenden eine Konkretisierung der Darstellung dieser Fallstricke, anhand derer es dem Prüfer ermöglicht wird, eine verfahrensunabhängige Vermeidungsstrategie bei Verwendung seiner professionellen Einschätzung zur Determinierung subjektiver Inputs bei Auswahlprüfungen zu entwickeln.

3.3.1 Anwendung ineffektiver Verfahren

Nicht alle Auswahlverfahren eignen sich zur Prüfung sämtlicher Prüffelder. So bieten sich statistische Stichprobenverfahren an, wenn homogene Massenvorgänge ohne geeignete Vorinformationen vorliegen, was jedoch in der Praxis selten der Fall sein dürfte[56]. Sie bieten sich zudem an, wenn das interne Kontrollsystem effektiv ist, da der notwendige Stichprobenumfang dann tendenziell geringer ausfallen kann. Besteht ein erhöhtes Kontrollrisiko, kann die gezielte Auswahl risikobehafteter Elemente jedoch zweckmäßiger sein. In einem solchen Fall dürfte sich zudem die für statistische Stichprobenverfahren vorab zu quantifizierende Fehlererwartung problematisch darstellen.

[55] Vgl. *Baumeister/Oldewurtel*, a.a.O. 2016, S. 171.
[56] Vgl. dazu auch *Hömberg*, BFuP 3/1997, S. 248 f.

Umfragen unter Abschlussprüfern[57] zeigen, dass statistische Verfahren häufig keine Anwendung finden, weil diese augenscheinlich die Verwendung des Prüfer-Knowhows einschränken[58]. Dabei ist jedoch zu beachten, dass auch bei Nutzung statistischer Methoden subjektive Entscheidungen getroffen werden müssen. Dazu gehören neben der Entscheidung zur Wahl der Methode an sich (Ziehungsverfahren, Art der Hochrechnung) auch notwendige Inputs zu erwarteten Eigenschaften der Population (Fehlererwartung, Fehlertoleranz) sowie zur Einschätzung der Zuverlässigkeit weiterer Prüfungshandlungen mit dem gleichen Prüfziel (notwendige Prüfungssicherheit, Konfidenzniveau).

Im Falle der Anwendung statistischer Verfahren wird häufig das Monetary Unit Sampling (MUS) das Verfahren der Wahl sein, welches jedoch ebenfalls mit gebotener Vorsicht anzuwenden ist. So ist der Grundsatz der Wesentlichkeit in zweierlei Maße zu interpretieren, d.h. der Prüfer hat stets auch das Risiko von Unterbewertungen zu würdigen. Insbesondere bei Anreizen zur Reduzierung des Ergebnisses (z.B. in Familienunternehmen oder Non-Profit-Organisationen), bei Prüfungen von Posten der Passivseite und bei Prüfungen auf Vollständigkeit kann aufgrund der überproportionalen Betrachtung höherwertiger Elemente nicht auf das MUS vertraut werden. In jenen Fällen bieten sich andere Prüfungshandlungen an, insbesondere physische Inventurbeobachtungen, die Prüfung der Periodenabgrenzung bestimmter Posten sowie analytische Prüfungshandlungen, die grundsätzlich sowohl für mögliche Über- als auch Unterbewertungen sensibilisieren und Indizien für mangelnde Vollständigkeit liefern können. Da die einfache und geschichtete Mittelwertschätzung komplexer in der Anwendung sind und regelmäßig höhere Stichprobenumfänge als das MUS benötigen, wird bei Prüfungen von Abschlussposten mit der Erwartung von sowohl Über- als auch Unterbewertungen einer bewussten Auswahl häufig Vorrang gegeben werden, auch wenn sich der Prüfer im Zweifel einer weniger fundierten Exkulpationsmöglichkeit aussetzt.

Folgende Beispiele fassen einige Auswahlprüfungen nach Stand der neuen Prüfungsstandards zusammen:

- Der Prüfer prüft einen Teil der Verbindlichkeiten aus Lieferungen und Leistungen. Die Auswahl erfolgt zufällig, ein Punktschätzer wird anhand der einfachen Mittelwertschätzung ermittelt. Der Stichprobenumfang wird durch den Prüfer

[57] Vgl. *Giezek*, Monetary Unit Sampling, Wiesbaden 2011, S. 184.
[58] Vgl. *Mochty*, a.a.O. 2012, S. 76.

genau wie im Vorjahr auf 40 festgelegt. → Es handelt sich mangels Berücksichtigung des Stichprobenrisikos um eine **nichtstatistische Stichprobe**.

- Im Rahmen der Prüfung der Periodenabgrenzung untersucht der Prüfer gezielt stichtagsnahe Transaktionen mit wesentlichen Beträgen. Er prüft diese in vollem Umfang anhand der entsprechenden Dokumente zum Gefahrenübergang. → Es handelt sich um eine eigens abgegrenzte Population, **eine Stichprobe wurde nicht gezogen**. Es kann keine Aussage über andere als die geprüften Elemente gemacht werden.
- Der Prüfer prüft sämtliche Forderungen gegen nahestehende Personen und versendet Saldenbestätigungen an die zehn Kunden mit dem höchsten Saldo der Forderungen aus Lieferungen und Leistungen. Diese machen 85 % des gesamten Forderungssaldos aus. Eine Hochrechnung auf die Grundgesamtheit findet nicht statt. → Der Prüfer hat **an keiner Stelle eine Stichprobe verwendet**.
- Anhand einer mathematischen Formel wird der Umfang für die Prüfung der Vorräte im Rahmen einer Inventurbeobachtung festgelegt. Die Auswahl der zu prüfenden Erzeugnisse erfolgte gezielt durch den Prüfer. Die festgestellten Fehler wurden auf die Grundgesamtheit hochgerechnet. → Hierbei handelt es sich nicht um eine Stichprobe, sondern um eine **bewusste Auswahl**. Eine Projektion auf die Grundgesamtheit hätte nicht stattfinden dürfen.

3.3.2 Stichprobenrisiko

Neben dem Nicht-Stichprobenrisiko[59], welches durch Anwendung ungeeigneter Prüfungshandlungen oder die Fehlinterpretation von Prüfungsnachweisen sowie das Nichterkennen von falschen Angaben oder Kontrollabweichungen bestehen kann, ist der Abschlussprüfer stets einem Stichprobenrisiko ausgesetzt, welches sich aufgrund des nicht vollständigen Einbezugs der Grundgesamtheit zwangsläufig ergibt. Das Stichprobenrisiko steht in wechselseitigem Zusammenhang mit dem Stichprobenumfang und spiegelt dabei das Risiko wider, dass der Prüfer aufgrund der Auswahlprüfung zu einem anderen Prüfungsurteil gelangt, als es bei einer Vollprüfung der Fall gewesen wäre[60]. Unterteilt wird es in das Risiko der irrtümlichen Ablehnung eines Prüffelds, obwohl keine wesentlichen Mängel vorliegen (*α-Risiko*) sowie das bedeutsamere Risiko der irrtümlichen

[59] Vgl. IDW EPS 310, Tz. 3d), Tz. A1.
[60] Vgl. IDW EPS 310, Tz. 3c), Tz. 9.

Annahme eines Prüffelds, obwohl wesentliche Mängel vorliegen (*β-Risiko*). Im Rahmen der professionellen Ermessensentscheidung bei der Verfahrenswahl zur Stichprobenuntersuchung muss dafür Sorge getragen werden, dass das Stichprobenrisiko angemessen berücksichtigt werden kann; eine quantitative Kontrolle des Stichprobenrisikos erlauben jedoch ausschließlich statistische Stichprobenverfahren auf Basis der Zufallsauswahl[61]. Nichtstatistische Verfahren können eine solche Quantifizierung im Rahmen des Repräsentationsschlusses nicht leisten, dennoch muss der Prüfer eine Berücksichtigung insbesondere des β-Risikos gewährleisten. Dabei ist zu berücksichtigen, dass eine Unterschreitung des rechnerisch notwendigen Stichprobenumfangs zu einem erheblichen Anstieg des β-Risikos führt[62].

Bei statistischen Verfahren unterstützt Prüfungssoftware die Beachtung des Stichprobenrisikos. Es stellt sich daher die Frage, wie eine angemessene Berücksichtigung auch bei anderen Verfahren sichergestellt werden kann. Den Prüfer wird bei Vermutung von Überbewertungen vor allem die obere Fehlerintensität interessieren. Diese gibt diejenige Abweichung an, die unter Berücksichtigung des akzeptierten Prüfungsrisikos maximal zu erwarten ist. Liegt dieser Wert oberhalb der festgelegten Toleranzwesentlichkeit, läuft der Prüfer Gefahr, ein Prüffeld mit wesentlichen Fehlern als ordnungsmäßig zu akzeptieren. Es wird dabei ein Intervall um den exakten hochgerechneten Wert für den wahren enthaltenen Fehler im Prüffeld (Punktschätzer) gelegt, um der inhärenten Unsicherheit der Auswahlprüfung Ausdruck zu verleihen. In der Praxis ist es daher verheerend, lediglich diesen Punktschätzer mit einer festgelegten Toleranzgrenze zu vergleichen, da die aufgrund der nicht vollständigen Untersuchung der Grundgesamtheit stets präsente Wahrscheinlichkeit, dass der wahre Fehler diesen Wert übersteigt, unberücksichtigt bleibt. Generell gilt in einem solchen Fall, dass es sich empfiehlt, die Stichprobe zu erweitern oder weitere Prüfungshandlungen durchzuführen, je näher der geschätzte Fehler dem tolerierbaren Fehler kommt[63]. Mangels Bestimmbarkeit konkreter Kosten-/Nutzenfunktionen ist jedoch insbesondere bei nichtstatistischen Stichprobenprüfungen lediglich eine approximative Annäherung an den optimalen Prüfungsumfang möglich.

[61] Vgl. *Weinand/Wolz,* WP Praxis 4/2012, S. 69, 74.
[62] Vgl. *Peek/Neter/Warren*, Auditing-J Pract Th 2/1991, S. 33-48.
[63] Vgl. *Mauldin/Wolfe*, CAR 3/2014, S. 663 f., die zeigen, dass Abschlussprüfer dazu neigen, Stichproben auch im Falle der Erlangung nur unzureichender Prüfungsnachweise aus Zeitgründen nicht zu erweitern.

Praktisch ergibt sich bei Berücksichtigung des Stichprobenrisikos das Problem, dem Mandanten im Zweifel eine triftige Begründung für die Ausweitung der Prüfungshandlung zu geben oder eventuelle Korrekturbedarfe nahezulegen[64], da eine Zuordnung des hochgerechneten Fehlers zu einzelnen Posten des Prüffeldes regelmäßig nicht möglich sein wird.

Abbildung 3.1 veranschaulicht einige Prüfungssituationen[65], in denen das Stichprobenrisiko eine wesentliche Rolle spielt:

Fall 1: Das Konfidenzintervall um den Punktschätzer von T€ 25 für den wahren aggregierten Fehler im Prüffeld übersteigt weder die obere noch die untere Toleranzwesentlichkeit. Der Prüfer geht davon aus, dass das Prüffeld maximal um T€ 50 unterbewertet oder um T€ 100 überbewertet ist. Dieser Fall wird am ehesten bei vielen kleinen Fehlern auftreten, die sich entsprechend kompensieren und stellt für den Prüfer den trivialsten Fall ohne weitere Korrekturbedarfe dar. Bei einer Prüfung auf Überbewertungen ist diese Konstellation nur denkbar, wenn keinerlei Fehler gefunden wurden.

Fall 2: Die obere Fehlerintensität übersteigt die Toleranzwesentlichkeit und schließt damit den Fall ein, dass eine wesentliche Überbewertung vorliegt. Der Fall stellt insbesondere bei der Prüfung von Posten der Aktivseite eine in der Praxis häufig vorkommende Prüfungssituation dar. Eine Nachbuchung wäre in diesem Fall notwendig. Eine Erweiterung des Prüfungsumfangs ist möglich, wobei die Wahrscheinlichkeit hoch ist, das bisher erlangte Ergebnis lediglich zu bestätigen. Dabei würde die Präzision steigen und das Konfidenzintervall schmaler werden.

Fall 3: In dieser Konstellation liegt die untere Fehlerintensität nahe der negativen Toleranzwesentlichkeit, wobei das Konfidenzintervall schmal ist und damit von einer präzisen Schätzung ausgegangen werden kann. Insbesondere im Fall einer nichtstatistischen Stichprobe ist das entsprechende Risiko einer irrtümlichen Annahme des Prüffelds damit hoch. Der Prüfer sollte in diesem Fall erwägen, weitere Prüfungshandlungen durchzuführen, um dieses Risiko zu verringern.

[64] ISA 450; IDW PS 250 n.F.
[65] Vgl. ausführlich *Hitzig*, The CPA Journal 5/2004, S. 34 f.

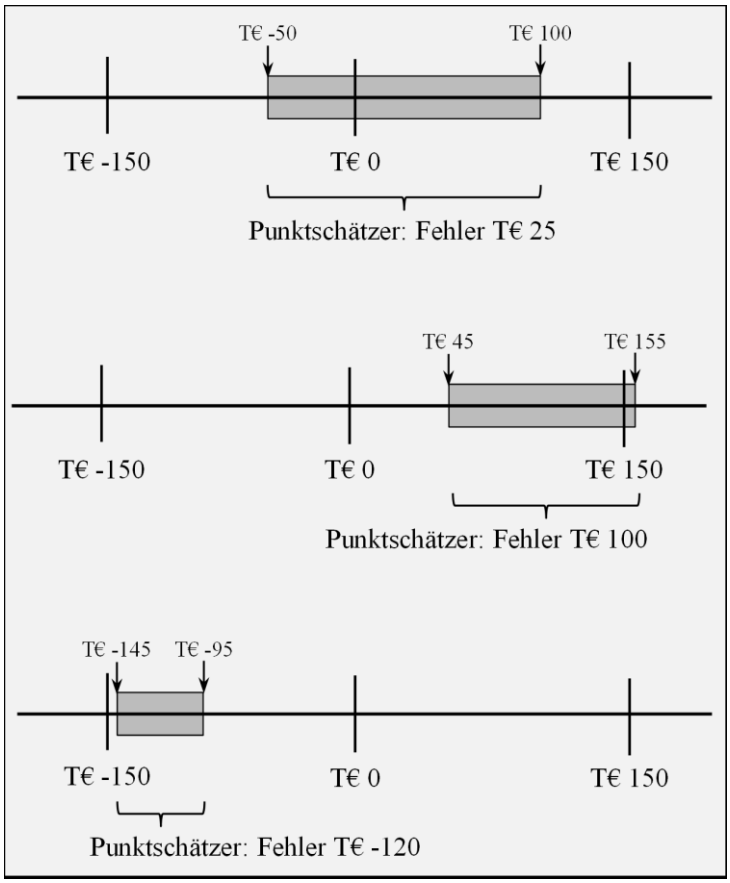


Abbildung 3.1: Berücksichtigung des Stichprobenrisikos mittels Fehlerintensitäten[66]

3.3.3 Die Isolierung von Fehlern

Grundsätzlich gilt für Stichproben, dass identifizierte Fehler auf die Grundgesamtheit hochgerechnet werden müssen, was bei Überschreitung des tolerierbaren Fehlers zu einem Korrekturbedarf führt. Praktisch gestaltet sich insbesondere die Würdigung entdeckter Fehler bei nichtstatistischen Stichproben diffizil, da die Projektion auf die Grundgesamtheit einer gewissen Subjektivität unterliegt. Der Prüfer kann bei qualitativer Berücksichtigung der Fehlerquellen dazu neigen, Fehler als Ausnahmen zu klassifizieren und von der Hochrechnung auszuschließen. Diese Problematik führt nicht selten dazu, dass Prüfer sich unbewusst einem hohen Risiko einer irrtümlichen Annahme des Prüffelds (β-Risiko) aussetzen. IDW EPS 310 geht im Gegensatz zur IDW St/HFA 1/1988 erstmals detaillierter auf die Behandlung solcher sog. Anomalien ein[67], bei denen es sich definitionsgemäß um „[…] eine falsche Angabe oder Kontrollabweichung, die nachweisbar nicht repräsentativ für falsche Angaben oder Kontrollabweichungen in einer Grundgesamtheit

[66] Eigene Darstellung in Anlehnung an ebd.
[67] Vgl. IDW EPS 310, Tz. 15.

ist“[68] handelt. Der Prüfungsstandard weist darauf hin, dass Anomalien nur in äußerst seltenen Fällen vorkommen und der Prüfer eine hohe Sicherheit darüber gewinnen muss, dass es sich um keinen repräsentativen Sachverhalt handelt. Die Wahrscheinlichkeit, einen isolierbaren Fehler in einer Grundgesamtheit von 1.000 Elementen bei einem Stichprobenumfang von 50 zu finden, beträgt im Mittel lediglich 5,3 %[69]. Der Prüfer wird demnach einen Teil der Grundgesamtheit isolieren müssen, um in diesem weitere Prüfungshandlungen durchzuführen, um die Annahme der Einzigartigkeit des Fehlers zu begründen. Abschlussprüfer können dazu neigen, durch den Ausschluss der Repräsentativität von Fehlern Wege zur Umgehung einer Erhöhung des Stichprobenumfangs zu finden, da diese eine Ausstrahlwirkung auf das beanspruchte Prüfungsbudget und das erreichbare Prüfungsurteil haben kann[70]. Grundsätzlich ist es jedoch auch im Falle der tatsächlichen Möglichkeit zur Isolierung von Fehlern zwingend geboten, die Anomalie im Gesamtumfang des gefundenen Fehlers im Prüffeld zu berücksichtigen[71], wobei eine Hochrechnung ausbleibt.

3.3.4 Stichprobenumfang

Hohe Stichprobenumfänge gelten als wesentlicher Nachteil statistischer Stichprobenverfahren[72]. Dass ein hoher Stichprobenumfang nicht erstrebenswert erscheint, verdeutlicht folgende Überlegung: Die Prüfungskosten stellen eine monoton steigende Funktion des Stichprobenumfangs dar. Rationales Verhalten unterstellt, wird der Abschlussprüfer seinen finanziellen Nutzen durch eine Minimierung des Stichprobenumfangs erreichen. In der Praxis wird aus diesem Grunde häufig mit einem nicht gerechtfertigten Stichprobenumfang gearbeitet, indem die Parameter der Inputgrößen zur Ermittlung des Stichprobenumfangs entsprechend angepasst werden („*working backward*“)[73]. Häufig findet zudem keine Revision des Stichprobenumfangs statt, selbst wenn Fehler aufgedeckt wurden[74].

Generell hat der Abschlussprüfer den Stichprobenumfang so festzulegen, dass das Stichprobenrisiko auf ein vertretbar niedriges Maß reduziert wird[75]. Wenngleich in den

[68] Vgl. IDW EPS 310, Tz. 7e).
[69] Vgl. *Hitzig*, The CPA Journal 9/2001, S. 50.
[70] Vgl. *Allen/Elder*, Auditing-J Pract Th 2/2005, S. 78-82.
[71] Vgl. dazu auch IDW PS 250 n.F., Tz. 24 ff.
[72] Vgl. *Giezek*, WPg 11/2014, S. 568.
[73] Vgl. empirisch z.B. *Messier/Kachelmeier/Jensen*, Auditing-J Pract Th 1/2001, S. 84.
[74] Vgl. *Mauldin/Wolfe*, CAR 3/2014, S. 663 f.
[75] IDW EPS 310, Tz. 9; ISA 530.7.

Anlagen zu IDW EPS 310 Hinweise gegeben werden, welche Faktoren Einfluss auf den Stichprobenumfang haben, bleibt die Nennung von Mindestumfängen aus. Aufgrund des geringen Grenznutzens von Einzelfallprüfungen bei gleichzeitig hohen Grenzkosten gerät der Prüfer in Versuchung, den Stichprobenumfang nicht ausreichend hoch zu wählen[76]. Zusammenfassend lassen sich folgende Faktoren mit Einfluss auf den Stichprobenumfang für Einzelfallprüfungen festhalten:

Tabelle 3.1: Einflussfaktoren auf den Stichprobenumfang[77]

Faktor	**Effekte, die zu geringerem Stichprobenumfang führen**	**Effekte, die zu höherem Stichprobenumfang führen**
Qualität des rechnungslegungsbezogenen internen Kontrollsystems	Höhere Verlässlichkeit bzw. geringeres Kontrollrisiko	Geringere Verlässlichkeit bzw. höheres Kontrollrisiko
Vertrauen in weitere aussagebezogene Prüfungshandlungen mit gleichem Prüfziel	Umfangreiche/effektive weitere Prüfungshandlungen mit gleichem Prüfziel (Zuverlässigkeit hoch)	Wenige/ineffektive weitere Prüfungshandlungen mit gleichem Prüfziel (Zuverlässigkeit mittel bis gering)
Prüffeld- oder postenbezogene Wesentlichkeit	Höhere Wesentlichkeitsgrenze	Geringere Wesentlichkeitsgrenze
Prüffeldbezogene Fehlererwartung	Niedrige erwartete Frequenz/Höhe	Mittlere bis hohe erwartete Frequenz/Höhe
Homogenität der Grundgesamtheit	Homogene Elemente (evtl. Bildung mehrerer Schichten)	Heterogene Elemente (evtl. keine/wenige Schichten)
Umfang der Grundgesamtheit	Geringer Effekt auf den zu verwendenden Stichprobenumfang, sofern Population nicht sehr klein	

Während diese Größen für eine statistische Stichprobe quantifiziert werden müssen, werden sie bei der bewussten Auswahl implizit berücksichtigt. Bei der nichtstatistischen Stichprobe helfen Tabellenwerke oder Arbeitshilfen, die jedoch weder das IDW noch das IFAC zur Verfügung stellen, was in der Praxis zu großen Spannweiten der verwendeten Stichprobenumfänge führen kann. Problematisch ist, dass die Wirkungsrichtungen einzelner Faktoren zwar bekannt sind, eine Aggregation jedoch i.d.R. nach prüferischem

[76] Vgl. *Elder/Akresh/Glover/Higgs/Liljegren*, Auditing-J Pract Th 2013, Suppl. 1, S. 100 f.
[77] Eigene Darstellung.

Ermessen erfolgt. Eine angemessene Dokumentation, um Einwendungen Dritter standhalten zu können, ist daher unumgänglich.

Der geplante Umfang durchzuführender Einzelfallprüfungen darf nicht ausschlaggebend dafür sein, welches Auswahlverfahren verwendet werden soll. Insbesondere kann die Verwendung nichtstatistischer Stichprobenverfahren nicht damit argumentiert werden, dass der benötigte Stichprobenumfang geringer ausfällt als bei der Nutzung statistischer Verfahren[78]. Gemäß IDW EPS 310, Tz. A10 soll der Umfang für statistische Stichprobenverfahren anhand einer „statistikbasierten Formel", für nichtstatistische Verfahren durch pflichtgemäßes Ermessen festgelegt werden. Da mithilfe statistischer Formeln ein optimaler Stichprobenumfang berechnet werden kann, scheint es unter ansonsten gleichen Umständen kaum möglich, dass der nach pflichtgemäßem Ermessen festgelegte Umfang geringer ausfällt.

In der Literatur wird empfohlen, dass für Grundgesamtheiten mit weniger als 200 Elementen nichtstatistische Stichprobenverfahren vorzuziehen sind[79]. In Abhängigkeit des Umfangs der Grundgesamtheit ist zu berücksichtigen, dass der relative Anteil der zu prüfenden Elemente im Prüffeld umso höher wird, je geringer der Umfang der Grundgesamtheit ausfällt. Bei Funktionstests von Kontrollinstanzen, die bspw. wöchentlich oder seltener durchgeführt werden, ist der Effekt der Stichprobendegression somit deutlich schwächer. Die Wahl eines (zu) geringen Stichprobenumfangs sorgt regelmäßig dafür, dass die erreichbare Präzision der Aussagen geringer ist, als vom Prüfer angenommen wird[80]. Bei der Wahl eines angemessenen Stichprobenumfangs sind in einem solchen Fall statistische Formeln oder entsprechende Tabellenwerke zu Rate zu ziehen[81]. Diese unterstützen die adäquate Berücksichtigung der den Stichprobenumfang determinierenden Faktoren wie Fehlererwartung, Fehlertoleranz sowie zu erreichende Aussagesicherheit. Da es sich bei diesen Größen jedoch ebenfalls um subjektive Einschätzungen des Prüfers handelt, sind auch die auf Basis solcher Hilfsmittel festgelegten Prüfungsumfänge stets kritisch zu hinterfragen und zwingend nachvollziehbar zu dokumentieren.

Ursächlich für die Verwendung zu geringer Stichprobenumfänge im Falle nichtstatistischer Stichprobenverfahren ist dabei häufig der sog. Ankereffekt (*Anchoring*), wonach

[78] Vgl. dazu empirisch z.B. *Swearingen/Hansen*, JABR 4/1990, S. 52.
[79] Vgl. *Jacoby/Hitzig*, The CPA Journal 12/2011, S. 34-36, wonach insbesondere bei Prüffeldern mit geringem Umfang die notwendigen Stichprobenumfänge häufig zu gering bemessen werden.
[80] Vgl. *Christensen/Elder/Glover*, Account Horiz 1/2015, S. 67, 70.
[81] Vgl. z.B. AICPA (Hrsg.), Audit Guide – Audit Sampling, New York 2014, S. Appendix A.11, C.1.

der Prüfer die Anzahl der zu prüfenden Elemente im pflichtgemäßen Ermessen anhand eines Initialwertes (z.B. Anzahl der geprüften Elemente im Vorjahr oder Umfang vergleichbarer Abschlussprüfungen) festlegt und in Abhängigkeit der vorliegenden Prüfungssituation lediglich eine Anpassung dieses Basiswertes vornimmt[82]. Dies führt dazu, dass die Umfänge und Spannweiten von Stichproben bei nichtstatistischen Stichproben und der bewussten Auswahl regelmäßig geringer sind als solche, die auf Basis statistischer Methoden ermittelt wurden. Eine formale Herleitung von Stichprobenumfängen führt demnach zu höheren Werten als solche, die der Prüfer im eigenen Ermessen als ausreichend erachtet. Eine solche unstrukturierte Heuristik ist in der Praxis in jedem Fall zu vermeiden, da einerseits ein Anstieg des β-Risikos bewusst in Kauf genommen wird[83] und eine solche Vorgehensweise die Möglichkeit der Exkulpation des Prüfers stark einschränkt.

3.4 Effizienzgewinn durch Schichtung

Insbesondere bei umfangreichen Prüffeldern mit streuenden Buchwerten kann eine Schichtung den Stichprobenumfang durch Homogenisierung der Grundgesamtheit senken[84], da sich in der Praxis häufig große Teile des gesamten Wertes eines Prüffeldes auf wenige Posten konzentrieren. Geschichtet werden kann nach prüferischem Ermessen oder auf Basis statistischer Formeln[85]. Regelmäßig wird der Prüfer eine Vollprüfung wertmäßig hoher Einzelposten anstreben und gleichzeitig aus der verbleibenden Grundgesamtheit eine Zufallsstichprobe ziehen.

Bei Verwendung des MUS sind solche Überlegungen obsolet, da methodisch bedingt bereits eine maximale Schichtung durchgeführt wird[86]. Eine Stichprobe aus zwei oder mehr Schichten macht eine geschichtete Hochrechnung entdeckter Fehler notwendig. Erst im Anschluss daran ist eine Gesamtaussage über das Prüffeld anhand eines Vergleichs mit der Toleranzwesentlichkeit möglich.

[82] Vgl. z.B. *Fay/Jenkins/Popova*, Managerial Auditing Journal 3/2015, S. 226-243; *Butler*, TAR 1/1986, S. 101-111.
[83] Vgl. *Butler*, a.a.O. 1986, S. 107.
[84] Vgl. IDW EPS 310, Anlage 1, Nr. 1-4; ISA 530, Appendix 1, Nr. 1-4.
[85] Vgl. *Schwartz*, The CPA Journal 2/1997, S. 57; *Heese/Braatsch*, a.a.O. 2013, S. 842.
[86] Vgl. *Giezek*, a.a.O. 2011, S. 35.

Typische Kriterien für eine sachgerechte Schichtung stellen bspw. dar[87]:

- Funktionen: Abgrenzung prozessualer Strukturen, die einen bestimmten Geschäftsbereich oder eine Gruppe von Geschäftsvorfällen definieren,
- Höhe des Betrages (z.B. Abgrenzung wesentlicher Forderungen oder Vorratsbestände),
- Abteilungen (Funktionale Zuordnung, um z.B. risikobehaftete Geschäftsbereiche stärker zu berücksichtigen),
- Aufträge (z.B. nicht regelmäßig abgesetzte Produkte),
- Kunden oder Lieferanten (z.B. Großabnehmer oder Auslandslieferanten),
- Verbundene Unternehmen und sonstige nahestehende Personen,
- Zeiträume (z.B. stichtagsnahe Sachverhalte oder Zeiträume, in denen funktionale Änderungen stattgefunden haben),
- Arbeitsgebiete bestimmter Mitarbeiter (auf Basis der im Laufe der Prüfung gewonnenen Risikoeinschätzung sowie bei Mitarbeiterwechseln oder -ausfällen).

3.5 Fazit und Ausblick

Die vorhergehenden Ausführungen haben anhand der Dreiteilung möglicher Auswahlverfahren in die bewusste Auswahl sowie statistische und nichtstatistische Stichproben sowie dem Aufzeigen der spezifischen Vor- und Nachteile dargelegt, dass der Abschlussprüfer eine weitreichende Freiheit hinsichtlich der Verfahrenswahl besitzt. Diese Feststellung ist nicht zwangsläufig unbefriedigend. In diesem Sinne ist bestehenden Darstellungen zu ISA 500[88] und ISA 530[89] beizupflichten, nachdem die Inhalte so interpretiert werden sollten, dass eine Verbindlichkeit statistischer Stichprobenverfahren zugunsten der Nutzung des Prüfer-Knowhows eben nicht erstrebenswert ist[90]. Eine weitreichende Änderung der Prüfungspraxis wird sich kaum ergeben. Es ist im Wesentlichen darauf zu achten, im Rahmen von Abschlussprüfungen, die bisher nicht bereits freiwillig nach den ISAs durchgeführt wurden, die Dokumentation hinsichtlich der Verfahrenswahl

[87] Vgl. hierzu und im Folgenden auch IDW (Hrsg.), WP-Handbuch, Band I, R, Düsseldorf 2012, Tz. 120, 128, S. 2433 f.
[88] Vgl. *Küster/Bernhardt*, WPg 23/2015.
[89] Vgl. *Heese/Braatsch*, a.a.O. 2013.
[90] Vgl. *ebd.*, S. 848.

anzupassen[91]. Es zeigt sich darüber hinaus jedoch, dass sich der Prüfer unabhängig vom gewählten Verfahren stets wesentlicher statistischer Zusammenhänge bewusst sein muss, ohne deren Kenntnisse er keine effektive Stichprobenprüfung durchführen kann. Für die erörterten Faktoren sensibilisiert IDW EPS 310 zwar, ohne dem Prüfer jedoch konkrete Handlungsempfehlungen bereitzustellen[92]. Vor dem Hintergrund der mannigfaltigen Möglichkeiten scheint die Entwicklung eines unterstützenden Instrumentariums sinnvoll, um den geneigten Prüfer für die Anwendung statistischer und nichtstatistischer Stichprobenverfahren zu sensibilisieren[93]. Nicht zuletzt bleibt auch die Frage nach dem Umfang etwaiger Ausbildungsbestrebungen im Rahmen der Berufsexamina, die bis dato eine Behandlung statistischer Verfahren im notwendigen Detailgrad vermissen lassen. Letzteres wird, zusammen mit den entsprechenden Haftungsbestimmungen im Berufsstand, keinen unwesentlichen Einfluss auf die mittelfristige Entwicklung des Einsatzes statistischer Auswahlverfahren haben. Eine Abkehr von der mehrheitlichen Anwendung der bewussten Auswahl sowie der nichtstatistischen Stichprobe scheint dabei unter gegebenen Umständen unwahrscheinlich.

[91] Vgl. zum Umfang der erforderlichen Dokumentation über IDW PS 460 hinaus auch *Mochty*, Nichtstatistische Stichproben, in: *Kirsch* (Hrsg.), Rechnungslegung und Wirtschaftsprüfung. Festschrift zum 70. Geburtstag von Jörg Baetge, Düsseldorf 2007, S. 1079 f.

[92] Vgl. dazu auch ebd., S. 1081, wonach das IDW im Falle einer nahezu wörtlichen Übersetzung des ISA 530 im Hinblick auf die erreichte Qualität handelsrechtlicher Jahresabschlussprüfungen deutliche Abstriche machen würde.

[93] Vgl. *Bonner/Libby/Nelson*, TAR 2/1996, S. 221-240 im Allgemeinen sowie *Messier/Kachelmeier/Jensen*, a.a.O. 2001, S. 81-96 im Speziellen, die zeigen, dass Leitfäden zur Wahl des Stichprobenumfangs nutzenstiftend sind.

4 Prüfungsnachweise mit Hilfe von Auswahlprüfungen – Empirische Analyse unter Berücksichtigung von IDW PS 300 n.F. und IDW PS 310

4.1 Publikationsdetails

Zusammenfassung: Mit IDW PS 300 n.F. und IDW PS 310 werden ISA 500 (Erlangung von Prüfungsnachweisen) und ISA 530 (repräsentative Stichproben) in nationale Prüfungsstandards transformiert. Damit erreichen die Vorgaben der IFAC auch den Mittelstand und Abschlussprüfungen, die bisher nicht freiwillig diesen Verlautbarungen folgten. Der vorliegende Beitrag zeigt anhand einer empirischen Auswertung, dass wesentliche Inhalte der überarbeiteten Prüfungsstandards bereits heute berufsübliche Praxis sind.

Koautoren: Christoph Oldewurtel, Prof. Dr. Christiane Pott, Martin Weinand.

Stichwörter: Auswahlverfahren, Stichprobe, Bewusste Auswahl, IDW PS 310, Prüfungsnachweis, IDW PS 300 n.F., HFA 1/1988, ISA 500, ISA 530.

Publikationsstatus: Veröffentlicht in: *WPg Die Wirtschaftsprüfung,* 71 (2): 73–82. Eine frühere Version des Artikels wurde im Dezember 2016 auf der Capital Market Based Accounting Research Conference, Münster, vorgestellt.

4.2 Einleitung

In Zeiten von Big Data sehen sich Abschlussprüfer aufgrund der für die Abschlussprüfung stetig größeren zu berücksichtigenden Datenmengen mehr denn je der Problematik der Auswahl einzuholender Prüfungsnachweise gegenüber. *IDW PS 300 n.F.*[94] und *IDW PS 310*[95] bieten dem Abschlussprüfer Vorgaben zur Erlangung von Prüfungsnachweisen sowie zu repräsentativen Auswahlverfahren (Stichproben) und stellen einen Gleichklang zu ISA 500[96] und ISA 530 her.[97] *IDW PS 310* ersetzt die für Auswahlprüfungen bisher gültige *Stellungnahme HFA 1/1988*[98] und sensibilisiert für potenzielle Konflikte aufgrund der Fehlinterpretation mathematisch-statistischer Gesetzmäßigkeiten. Dies gilt vor allem für die Anwendung der in der Berufspraxis häufig genutzten nicht-statistischen Stichprobenverfahren.

4.3 Anwendungsbereiche von IDW PS 300 n.F. und IDW PS 310

IDW PS 300 n.F. gilt grundsätzlich für die Einholung von Prüfungsnachweisen und damit für jede Abschlussprüfung. Werden Prüfungsnachweise zu einem Prüffeld nicht in vollem Umfang eingeholt, während das Prüfungsurteil jedoch zum gesamten Prüfgebiet erfolgen soll (Auswahlprüfung), sind künftig stets die entsprechenden Vorschriften gemäß *IDW PS 300 n.F.*, Tz. 11, und ggf. gemäß *IDW PS 310* anzuwenden. Auswahlprüfungen kommen vor allem dann zum Tragen, wenn die Zahl der Elemente im Prüffeld groß ist,[99] wenige oder keine Informationen zu einem erhöhten Risiko spezifischer Elemente vorliegen und die technischen Voraussetzungen zur Datenerhebung beim Mandanten gegeben sind. Der Abschlussprüfer bedient sich der Auswahlprüfung sowohl im Rahmen von Funktionsprüfungen des internen Kontrollsystems als auch bei der Durchführung aussagebezogener Einzelfallprüfungen (*IDW PS 300 n.F.*, Tz. 11).[100]

---

[94] Vgl. *IDW Prüfungsstandard: Prüfungsnachweise im Rahmen der Abschlussprüfung (IDW PS 300 n.F.)* (Stand: 14.06.2016); vgl. zu *IDW PS 300 n.F.* allgemein Küster/Bernhardt, WPg 2015, S. 1212.

[95] *IDW Prüfungsstandard: Repräsentative Auswahlverfahren (Stichproben) in der Abschlussprüfung (IDW PS 310)* (Stand: 14.06.2016).

[96] ISA 500 „Audit Evidence", in: IFAC (Hrsg.), 2015 Handbook of International Quality Control, Auditing, Review, Other Assurance and Related Services Pronouncements, New York 2015, Vol. I; zu einer deutschen Übersetzung der ISA (Stand: April 2010) siehe IDW Textausgabe, International Standards on Auditing (ISAs), Düsseldorf 2011.

[97] ISA 530: „Audit Sampling", in: IFAC (Hrsg.), a.a.O. (Fn. 3); zu ISA 530 allgemein Heese/Braatsch, WPg 2013, S. 841 ff.

[98] *HFA Stellungnahme: Zur Anwendung stichprobengestützter Prüfungsmethoden bei der Jahresabschlussprüfung (HFA 1/1988)*.

[99] Vgl. Griffin/Wright, Accounting Horizons 2015, S. 377.

[100] Vgl. detailliert IDW (Hrsg.), WP Handbuch 2012, Bd. I, 14. Aufl., Düsseldorf 2012, Kap. R, Tz. 119.

Erfolgen die Auswahl und ggf. die Auswertung unter Berücksichtigung mathematisch-statistischer Gesetzmäßigkeiten, handelt es sich um eine Stichprobe gemäß *IDW PS 310*, Tz. 7g). Dabei werden ausgewählte Elemente aus der Menge aller Einzelposten des Prüffelds (Grundgesamtheit) untersucht, um Aussagen über das gesamte Prüffeld treffen zu können (Inferenz). Da sich Kontrollabweichungen oder falsche Angaben in Massendaten in der Regel nicht allein durch Systemprüfungen und analytische Prüfungshandlungen hinreichend prüfen lassen,[101] fallen Prüfungshandlungen nahezu jeder Abschlussprüfung in den Anwendungsbereich von *IDW PS 300 n.F.* und *IDW PS 310*.[102]

4.4 Verfahren zur Erlangung von Prüfungsnachweisen

4.4.1 Abgrenzung wirksamer Verfahren

Bei der Planung und Durchführung von Funktions- und Einzelfallprüfungen sind grundsätzlich wirksame Verfahren zur Auswahl der zu prüfenden Elemente festzulegen, um ausreichende und angemessene Prüfungsnachweise zu erlangen (*IDW PS 300 n.F.*, Tz. 7). Dies ist einerseits durch eine Vollerhebung erreichbar (*IDW PS 300 n.F.*, Tz. A49). Andererseits stellt eine vollkommene Urteilssicherheit im Sinne des risikoorientierten Prüfungsansatzes keine Notwendigkeit dar, sodass die Einholung von weniger als 100% der möglichen Prüfungsnachweise für die Urteilsbildung über ein Prüffeld hinreichend ist.[103] Zur Erlangung angemessener Prüfungsnachweise stehen dem Prüfer grundsätzlich die in Abbildung 4.1 gezeigten Verfahren zur Verfügung (*IDW PS 300 n.F.*, Tz. A48).

[101] Vgl. auch *IDW Prüfungsstandard: Feststellung und Beurteilung von Fehlerrisiken und Reaktionen des Abschlussprüfers auf die beurteilten Fehlerrisiken (IDW PS 261 n.F.)* (Stand: 14.06.2016).

[102] *IDW PS 312*, Tz. 12, verpflichtet den Abschlussprüfer bei Vorliegen bedeutsamer Risiken gleichwohl zur Durchführung aussagebezogener Prüfungshandlungen; vgl. *IDW Prüfungsstandard: Analytische Prüfungshandlungen (IDW PS 312)* (Stand: 13.03.2013).

[103] In derartige Wirtschaftlichkeitsbetrachtungen einzubeziehen sind jedoch Kosten, die als „Rüstkosten" der Auswahlprüfung entstehen.

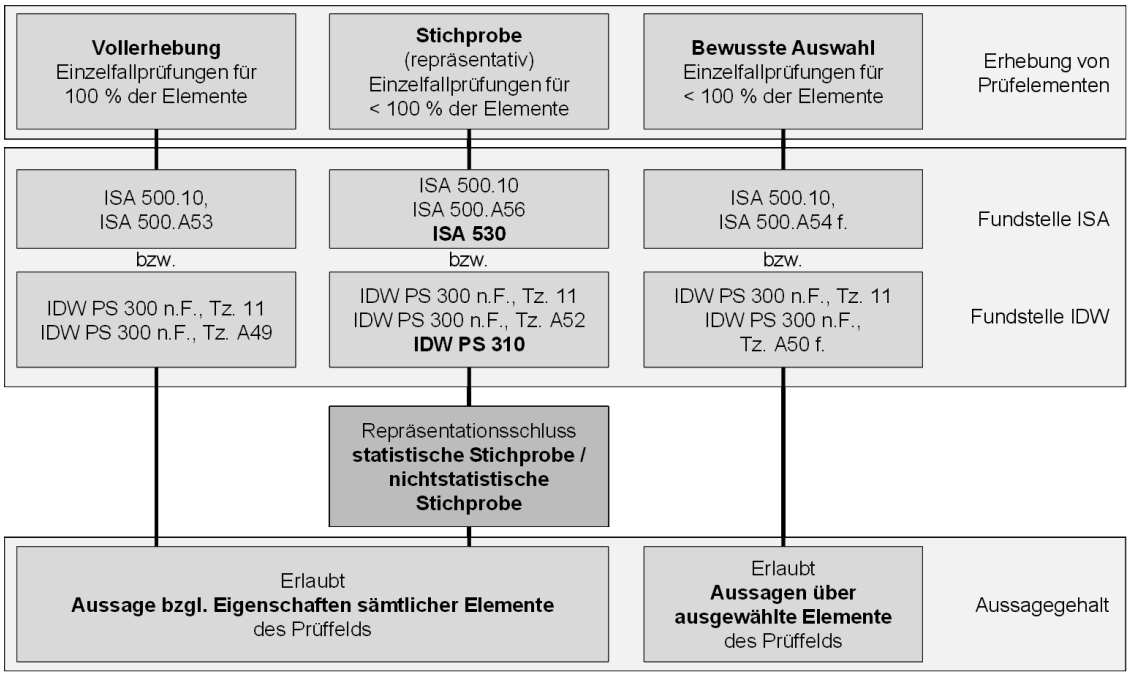


Abbildung 4.1: Verfahren zur Erlangung angemessener Prüfungsnachweise

4.4.2 Vollerhebung versus bewusste Auswahl und repräsentative Stichprobe

Eine Vollerhebung kann zielführend sein, wenn das zu beurteilende Prüffeld ein bedeutsames Risiko darstellt und/oder nur eine geringe Zahl von Elementen enthält. Sie kann zudem effizient sein, wenn eine IT-gestützte Massendatenanalyse unter Berücksichtigung aller Elemente der Grundgesamtheit möglich ist.[104]

Liegen dem Prüfer Informationen über das Prüffeld vor, anhand derer er eine bewusste Auswahl nach prüferischem Ermessen (professional judgment) rechtfertigen kann, wird er es im Sinne des risikoorientierten Prüfungsansatzes[105] präferieren, Elemente mit einer hohen Fehleranfälligkeit auszuwählen (*IDW PS 300 n.F.*, Tz. A50).[106] Der Prüfer kann diese Informationen nutzen, um das verbleibende Prüffeld im Rahmen des prüferischen Ermessens unter Risiko- und Wesentlichkeitsgesichtspunkten als ordnungsmäßig anzunehmen.

Erfolgen die Auswahl und ggf. die Auswertung von Prüfungsnachweisen auf der Basis mathematisch-statistischer Gesetzmäßigkeiten, handelt es sich um eine Stichprobe im

[104] Vgl. *IDW PS 300 n.F.*, Tz. A49; Heese/Braatsch, WPg 2013, S. 842; zur Massendatenanalyse auch *IDW Prüfungshinweis: Einsatz von Datenanalysen im Rahmen der Abschlussprüfung (IDW PH 9.330.3)* (Stand: 15.10.2010).
[105] Vgl. *IDW PS 261 n.F.*, Tz. 5 ff.
[106] Vgl. kritisch Mochty, Nichtstatistische Stichproben, in: Kirsch/Thiele (Hrsg.), FS Baetge, Düsseldorf 2007, S. 1060.

Sinne von *IDW PS 310*.[107] Dabei wird in Abhängigkeit von der Art der Durchführung zwischen statistischen und nicht-statistischen Stichproben unterschieden (*IDW PS 310*, Tz. 7g)). Stichprobenelemente werden im Gegensatz zur bewussten Auswahl so ermittelt, dass der Prüfer keinen Einfluss auf die Auswahlwahrscheinlichkeiten hat (Zufallsauswahl). Sie werden zu statistisch verwertbaren Zufallsvariablen und in der Regel mittels EDV-Unterstützung per echter oder unechter Zufallsauswahl ermittelt.[108]

Da die klassische statistische Methodenlehre keine für die Abschlussprüfung universell einsetzbaren Methoden bereithält, wurden Verfahren wie das Monetary Unit Sampling (MUS) entwickelt.[109] *IDW PS 310*, Tz. A8, betont die Möglichkeit einer solchen wertproportionalen Stichprobenauswahl.[110]

4.4.3 Kombination von bewusster und zufälliger Auswahl

In der Praxis kommt es häufig zu einer Kombination von Auswahlmethoden. Dieses Vorgehen gilt im Sinne der Prüfungsstandards explizit als angemessen (*IDW PS 300 n.F.*, Tz. A48), da einerseits dem Vorwissen des Prüfers Rechnung getragen wird, andererseits seine Unabhängigkeit gewahrt bleibt. Abbildung 4.2 verdeutlicht eine solche Kombination im Falle der bewussten und zufälligen Auswahl. Werden einige Elemente eines Prüffelds mittels bewusster Auswahl isoliert betrachtet, während die verbleibenden Elemente der Grundgesamtheit auf der Basis einer Zufallsauswahl gesondert gewürdigt werden, handelt es sich um eine Schichtung der Grundgesamtheit. Das Schichtungsmerkmal wird in der Praxis regelmäßig der Buchwert (Ist-Wert) der Untersuchungsobjekte sein. In diesem Fall handelt es sich um die Kombination einer bewussten Auswahl gemäß *IDW PS 300 n.F.*, Tz. A50, mit einer Stichprobenprüfung im Sinne von *IDW PS 310*.

[107] Vgl. Abschnitt 4.5.2.
[108] Vgl. für die Darstellung verfügbarer Ziehungsverfahren z.B. Weinand/Wolz, WP Praxis 2012, S. 69 f.
[109] Vgl. Leslie/Teitlebaum/Anderson, Dollar Unit Sampling, Toronto 1979.
[110] Bei dieser Methode wird davon ausgegangen, dass die erwarteten Abweichungen mit den Buchwerten korrelieren, was vor allem auf vermutete Überbewertungen und damit Aktivposten und Erträge zutrifft (*IDW PS 310*, Anlage 1, Nr. 5).

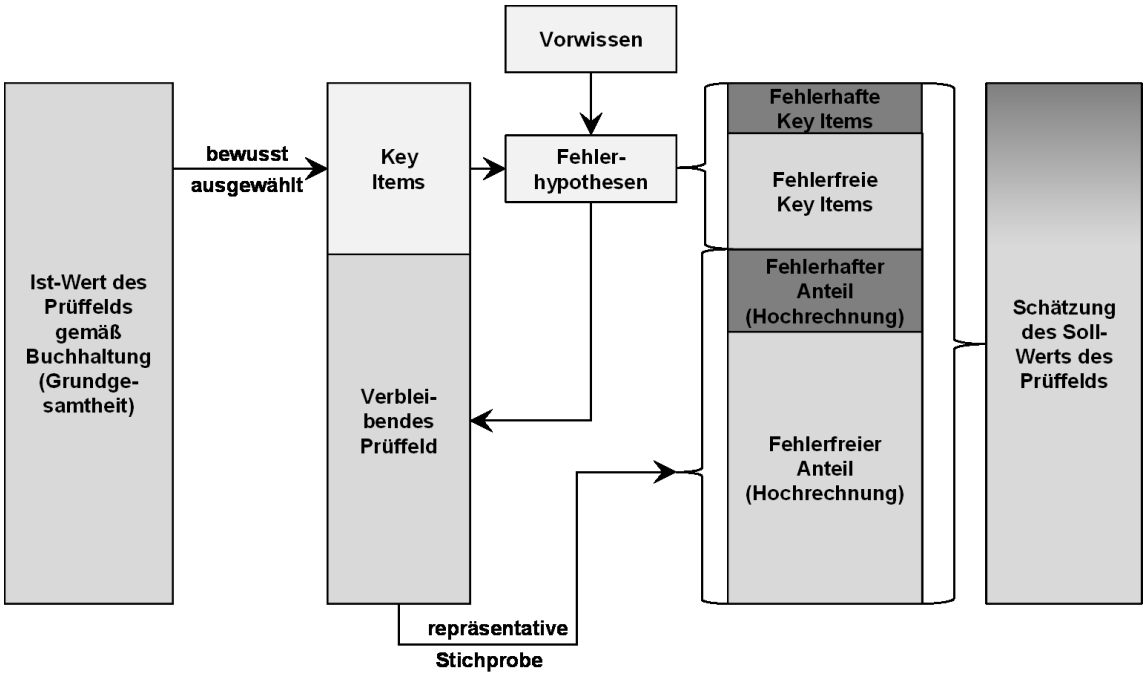


Abbildung 4.2: Kombination aus bewusster Auswahl und Stichprobe

Insgesamt ergibt sich durch verschiedene Kombinationsmöglichkeiten eine große Bandbreite an verwendbaren Methoden im Sinne von *IDW PS 300 n.F.* und *IDW PS* 310. Tabelle 4.1 veranschaulicht dies anhand der beiden Ausprägungen einer repräsentativen Stichprobe sowie bewusst ausgewählter Elemente.

Tabelle 4.1: Eigenschaften der bewussten Auswahl und der repräsentativen Stichprobe

	Bewusste Auswahl nach IDW PS 300 n.F.	…	**Repräsentative Stichprobe nach IDW PS 310**
Gegen-stand	Komplexer kognitiver Urteilsbildungsprozess	…	Stichprobentechnik nach mathematisch-statistischen Grundsätzen
Auswahl	Nach pflichtgemäßem Ermessen (selektive Auswahl)	…	Unkontrolliert (echte oder unechte Zufallsauswahl), repräsentativ
Ergebnis	Qualitatives Urteil über das Prüffeld Annahme/Ablehnung nach pflichtgemäßem Ermessen	…	Quantifizierbare Aussage über das Prüffeld (statistische Grundgesamtheit) Gegebenenfalls qualitative Beurteilung des Stichprobenrisikos

4.5 Ausgewählte Regelungsinhalte von IDW PS 300 n.F. und IDW PS 310

4.5.1 Stellenwert der bewussten Auswahl

Bisher galt die bewusste Auswahl gemäß *Stellungnahme HFA 1/1988*, Abschnitt C.I., als Stichprobe. Durch die ausschließliche Regelung von Stichproben in *IDW PS 310* wird der bewussten Auswahl dieser Status aberkannt. Gleichwohl ändert sich durch diese Verlagerung nichts am möglichen Aussagegehalt einer bewussten Auswahl. Weiterhin gilt, dass eine ausschließlich selektive Auswahl von Prüfobjekten aus der Grundgesamtheit dem Prüfer keine Aussage über die Elemente außerhalb dieser ausgewählten Menge erlaubt. Eine quantitative Hochrechnung verbietet sich entsprechend.[111] Hiervon unberührt kann in vielen Fällen eine bewusste Auswahl zweckmäßig sein, da sie in hohem Maße die Nutzung von Vorinformationen und Prüferwissen erlaubt und somit den zuvor im Rahmen der Prüfungsplanung identifizierten mandatsspezifischen Risiken auf der Ebene einzelner Prüffelder in besonderem Maße Rechnung getragen werden kann. Wirtschaftlich betrachtet wird sie regelmäßig sinnvoller einzusetzen sein als ein formales statistisches Erhebungsverfahren. Es ist jedoch darauf zu achten, die Dokumentation in den Arbeitspapieren den neuen Gegebenheiten entsprechend anzupassen.[112]

4.5.2 Abgrenzung statistischer und nicht-statistischer Stichprobenverfahren

Im Gleichklang mit den ISA stellt *IDW PS 300 n.F.*, Tz. A51, zur bewussten Auswahl wiederholend klar, dass eine selektive Untersuchung von Elementen häufig ein wirtschaftliches Mittel zur Erlangung von Prüfungsnachweisen ist, dabei jedoch keine Stichprobenprüfung darstellt. Um ein statistisches Stichprobenverfahren handelt es sich allein dann, wenn folgende Bedingungen kumulativ erfüllt werden (*IDW PS 310*, Tz. 7g)):

- zufallsgesteuerte Auswahl der zu prüfenden Elemente,[113]
- Anwendung der Wahrscheinlichkeitstheorie zur Auswertung der Stichprobenergebnisse (samt Bewertung des Stichprobenrisikos).

Ein in der Praxis häufig anzutreffender, vom Prüfer intuitiv festgelegter Prüfungsumfang stellt – selbst bei Verwendung der Zufallsauswahl – hingegen eine nicht-statistische

[111] Vgl. Burgstahler/Glover/Jiambalvo, Auditing 2000, S. 82.
[112] Inwiefern dies auch auf den Bestätigungsvermerk aufgrund des dortigen Hinweises auf die überwiegende Verwendung von Stichprobenverfahren ausstrahlt, wird hier nicht weiter untersucht.
[113] Vgl. für eine Darstellung berufsüblicher Methoden der Zufallsauswahl Heese/Braatsch, WPg 2013, S. 845 f.; Weinand/Wolz, WP Praxis 2012 S. 69.

Stichprobe dar. Statistische Stichprobenverfahren gemäß *IDW PS 310*, Tz. A11, sind entsprechend (nur) solche, die jede Art von systematischer Verzerrung vermeiden und auf diese Weise einen Repräsentationsschluss zulassen.

4.5.3 Berücksichtigung des Stichprobenrisikos und Ermittlung des notwendigen Stichprobenumfangs

Der Umfang aussagebezogener Prüfungshandlungen kann in Form des geforderten Sicherheitsgrads mit Auswirkung auf den benötigten Stichprobenumfang variiert werden.[114] Die unvollständige Prüfung der Grundgesamtheit führt zu einer Unsicherheit bezüglich der Präzision des Urteils (Stichprobenrisiko). Neben dem Nicht-Stichprobenrisiko,[115] welches unter anderem durch Anwendung ungeeigneter Prüfungshandlungen oder durch die Fehlinterpretation von Prüfungsnachweisen entsteht, ist der Abschlussprüfer diesem stets ausgesetzt. Es ist vor allem hinsichtlich der irrtümlichen Annahme eines Prüffelds kritisch und steht in wechselseitigem Zusammenhang zum benötigten Stichprobenumfang.[116]

Eine Berücksichtigung des Stichprobenrisikos ist quantitativ nur möglich, wenn ein Stichprobenverfahren verwendet wird, das die Bedingungen von *IDW PS 310*, Tz. 7g), erfüllt. In diesem Fall wird ein Konfidenzintervall um den Punktschätzer des Soll-Buchwerts des Prüffelds gebildet.[117] Ist das Stichprobenrisiko (zu) hoch, muss der Prüfungsumfang ggf. erweitert werden.[118]

Der Stichprobenumfang kann in Abhängigkeit vom Stichprobenrisiko sowie von weiteren den Prüfungsumfang determinierender Faktoren[119] entweder formal oder nach prüferischem Ermessen ermittelt werden (Tabelle 4.2). Generell hat der Abschlussprüfer einen Stichprobenumfang zu wählen, der ein vertretbares Maß des Stichprobenrisikos realisiert.[120]

---

[114] Vgl. *IDW PS 261 n.F.*, Tz. 82.
[115] Vgl. *IDW PS 310*, Tz. 3d), Tz. A1.
[116] Vgl. *IDW PS 310*, Tz. A10.
[117] Vgl. *IDW PS 310*, Tz. A20b; im Detail Baumeister/Oldewurtel, WP Praxis 2016, S. 201 f.
[118] Vgl. empirisch Mauldin/Wolfe, Contemporary Accounting Review 2014, S. 663 f., die für ein amerikanisches Sample zeigen, dass Stichproben auch im Falle unzureichender Prüfungsnachweise regelmäßig nicht erweitert werden.
[119] Vgl. *IDW PS 310*, Anlagen 1 und 2.
[120] Vgl. *IDW PS 310*, Tz. 9; dabei sollte er sich keinesfalls allein am Vorjahr orientieren; vgl. zum sog. „SALY"-Ansatz (Same-as-Last-Year) z.B. Fay/Jenkins/Popova, Managerial Auditing Journal 2015, S. 226.

Tabelle 4.2: Faktoren mit Einfluss auf den Stichprobenumfang

Einflussfaktor	Effekte, die zu geringerem Stichprobenumfang führen	Effekte, die zu höherem Stichprobenumfang führen
Zuverlässigkeit des rechnungslegungsbezogenen internen Kontrollsystems	Höhere Verlässlichkeit bzw. geringeres Kontrollrisiko	Geringere Verlässlichkeit bzw. höheres Kontrollrisiko
Vertrauen in weitere aussagebezogene Prüfungshandlungen mit gleichem Prüfziel	Umfangreiche/effektive weitere Prüfungshandlungen mit gleichem Prüfziel (Zuverlässigkeit hoch)	Wenige/ineffektive weitere Prüfungshandlungen mit gleichem Prüfziel (Zuverlässigkeit mittel bis gering)
Prüffeld- oder postenbezogene Wesentlichkeit	Höhere Wesentlichkeitsgrenze	Geringere Wesentlichkeitsgrenze
Prüffeldbezogene Fehlererwartung	Niedrige erwartete Fehlerfrequenz/-höhe	Mittlere bis hohe erwartete Fehlerfrequenz/-höhe
Homogenität der Grundgesamtheit	Homogene Elemente (geringe Streuung von Buchwerten und erwarteten falschen Darstellungen)	Heterogene Elemente (mittlere bis hohe Streuung von Buchwerten und erwarteten falschen Darstellungen)
Umfang der Grundgesamtheit	Geringer Effekt auf den zu verwendenden Stichprobenumfang (Stichprobendegression), sofern Umfang der Grundgesamtheit nicht sehr klein ist	

4.5.4 Umgang mit Anomalien und Fehlerisolierungen

Für Stichproben gilt grundsätzlich, dass identifizierte Fehler auf die Grundgesamtheit hochzurechnen sind. Der Prüfer kann bei der qualitativen Bewertung von Fehlerquellen dazu neigen, Fehler als Ausnahmen zu klassifizieren und von einer Hochrechnung auszuschließen (Fehlerisolierung). *IDW PS 310* geht im Gegensatz zur *Stellungnahme HFA 1/1988* auf den Umgang mit solchen „Anomalien" ein, bei denen es sich definitionsgemäß um Fehler handelt, die nachweisbar nicht repräsentativ für die Grundgesamtheit sind.[121] *IDW PS 310* beschränkt das Vorkommen von Anomalien dabei auf äußerst seltene Fälle.[122] Der Prüfer hat die Prüfungshandlungen ggf. so auszuweiten, dass eine abgrenzbare Teilmenge der Grundgesamtheit genauer untersucht wird.[123] Eine Anomalie ist in

[121] Vgl. *IDW PS 310*, Tz. 7e).
[122] Vgl. *IDW PS 310*, Tz. 15.
[123] Vgl. *IDW PS 310*, Tz. A15.

den Gesamtumfang des ermittelten Fehlers im Prüffeld aufzunehmen, jedoch unterbleibt eine Hochrechnung.[124]

4.5.5 Behandlung nicht prüfbarer Elemente

Ist eine Prüfungshandlung nicht durchführbar, muss der Prüfer ein Ersatzelement der geplanten Prüfungshandlung unterziehen. *IDW PS 310* unterscheidet zwei Szenarien, in denen die Durchführung von Prüfungshandlungen in erster Instanz nicht durchführbar ist (*IDW PS 310*, Tz. 12 f., Tz. A13 f.):

- Die Nicht-Anwendbarkeit einer Prüfungshandlung liegt vor, wenn das Untersuchungsmerkmal des Stichprobenelements nicht dem ursprünglichen unterstellten Informationsgehalt entspricht. Es ist ein Ersatzelement zu ziehen, auf das die Prüfungshandlung angewendet wird. Die Nicht-Anwendbarkeit stellt keinen Fehler dar und hat daher keine Auswirkungen auf die Hochrechnung.
- Die Unmöglichkeit einer Prüfungshandlung liegt vor, wenn weder primäre noch alternative Prüfungshandlungen an einem Stichprobenelement durchgeführt werden können. Die Unmöglichkeit stellt einen Fehler dar und wird entsprechend hochgerechnet.

4.6 Empirische Untersuchung

4.6.1 Untersuchungsgegenstand

Zur Erhebung des Status Quo der Anwendung und Ausgestaltung von Auswahlprüfungen in der Abschlussprüfungspraxis wurden Arbeitspapiere zu insgesamt 340 Prüffeldern aus 73 Engagements analysiert, deren Abschlussstichtage in den Zeitraum 31.12.2014 bis 30.04.2016 fallen. Bei der Auswahl wurde angestrebt, einen Querschnitt der typischen Mandantenstruktur mittelständischer sowie second-tier-Prüfungsgesellschaften abzubilden (Abbildung 4.3). Im Durchschnitt erwirtschafteten die geprüften Unternehmen Umsatzerlöse i.H. von 26,5 Mio. € und wiesen zum jeweiligen Stichtag eine Bilanzsumme in Höhe von 18,6 Mio. € auf. Von 73 Engagements handelt es sich bei 68 Unternehmen um Kapital- und Personenhandelsgesellschaften, davon acht Mandate gemäß § 319a HGB. Im Rahmen einer Pilotstichprobe wurden Referenzarbeitspapiere zu 15 Engagements

[124] Vgl. *IDW PS 310*, Tz. A17; im Weiteren auch *IDW PS 250 n.F.*, Tz. 24 ff.

gesichtet und Untersuchungsschwerpunkte festgelegt.[125] Die Auswertung erfolgte anonymisiert.

in T€	Bilanzsumme	Umsatzerlöse	Periodenergebnis
Maximum	959.242	3.160.145	57.918
Minium	587	0	-12.045
Arithmetisches Mittel	61.179	117.839	2.629
Median	18.616	26.468	515
25 %-Quartil	9.097	11.304	112
75 %-Quartil	62.188	84.263	1.974

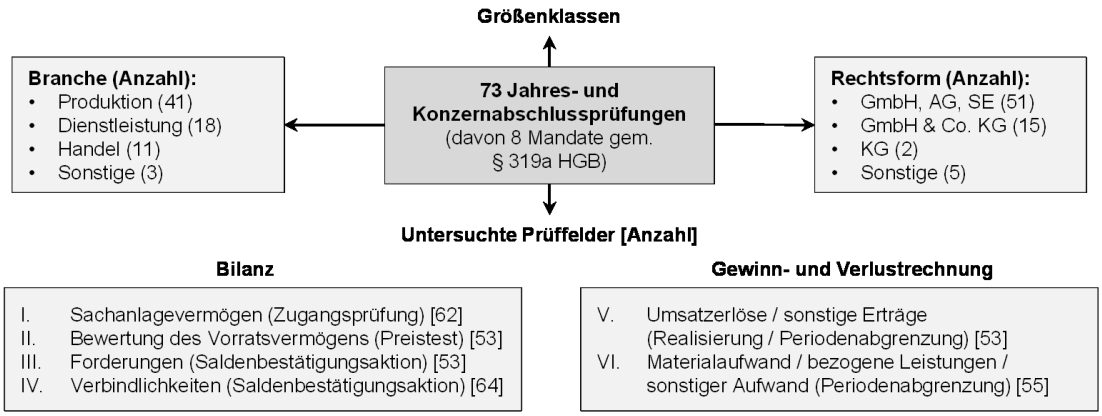


Abbildung 4.3: Untersuchungsgegenstand

4.6.2 Ergebnisse

4.6.2.1 Erhebungs- und Auswertungsverfahren

Abbildung 4.4 zeigt, dass etwa die Hälfte aller Fälle auf der bewussten Auswahl basiert (173 von 340 Prüffeldern). Die bewusste Auswahl ist vor allem bei Prüfungshandlungen im Anlagevermögen sowie bei der Prüfung passiver Bilanzposten bzw. von Aufwandsposten zu beobachten (109 von 181). In diesen Prüffeldern wird regelmäßig eine hohe Abdeckung des Postensaldos erreicht, was in nahezu allen Fällen mit der Fokussierung auf risikobehaftete Elemente begründet wird.

Demgegenüber ist das Methodenspektrum bei der Prüfung der Vorräte und Forderungen und bei der Prüfung von Erträgen ausgeglichen. Die bewusste Auswahl wird vor allem bei übersichtlichen Prüffeldern kleiner und mittelgroßer Unternehmen verwendet. In 101 von 159 Fällen (63,5%) überwiegt eine mindestens zum überwiegenden Teil verwendete echte Zufallsauswahl.

[125] Funktionsprüfungen des internen Kontrollsystems werden aufgrund der nahezu ausschließlichen Verwendung der einfachen Zufallsauswahl nicht betrachtet.

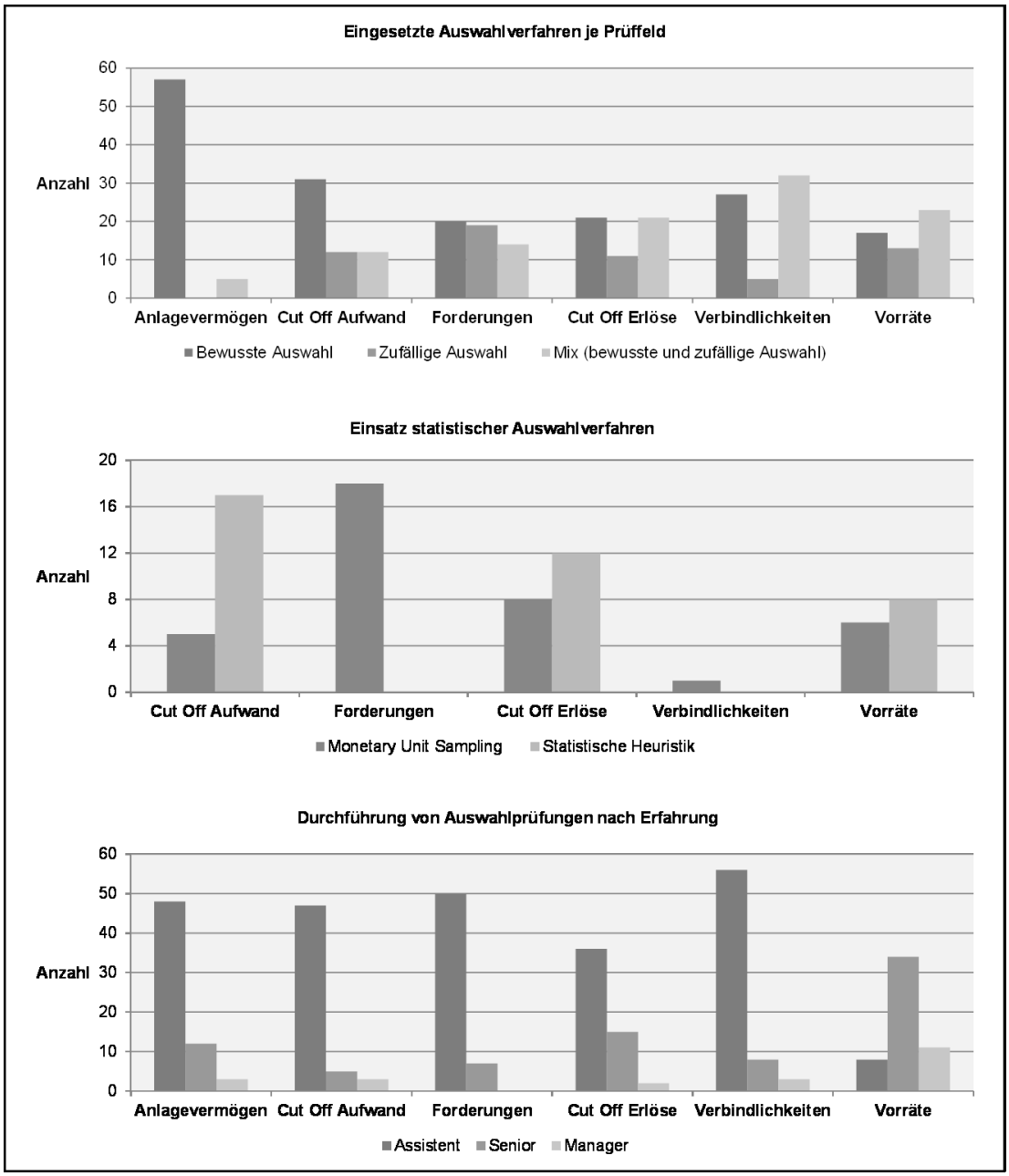


Abbildung 4.4: Einsatz von Auswahlverfahren in Abhängigkeit vom Prüffeld

Der Einsatz repräsentativer Auswahlverfahren gemäß *IDW PS 310* wird vorrangig bei der Prüfung der Vorräte, Forderungen sowie Posten der Gewinn- und Verlustrechnung realisiert. Dabei werden regelmäßig Heuristiken – z.B. die Bestimmung eines geeigneten Stichprobenumfangs anhand von Hilfstabellen – genutzt. Es zeigt sich auch, dass das MUS vor allem bei Posten mit Tendenz zur Überbewertung üblich ist.[126]

[126] Vgl. Giezek, WPg 2014, S. 569.

4.6.2.2 Prüfungsumfänge

Der Umfang von Prüfungshandlungen weist in Abhängigkeit vom verwendeten Erhebungsverfahren sowie vom Prüffeld eine hohe Varianz auf (Tabelle 4.3). Während bei der bewussten Auswahl im Mittel 15 Elemente geprüft werden, sind es bei Anwendung statistischer Heuristiken bereits 27 Elemente und beim Einsatz des MUS 31 Elemente (mit Oberschichtwerten 37 Elemente). Darüber hinaus sind die Prüfungsumfänge bei der Prüfung von Posten der Gewinn- und Verlustrechnung sowie der Vorräte tendenziell höher. Wenngleich eine Korrelation des Prüfungsumfangs mit dem Umfang der Grundgesamtheit festzustellen ist, kann dies – konform mit der Erläuterung des Effekts der sogenannten Stichprobendegression in *IDW PS 310*, Anlage 3 – nicht der alleinige Grund für diese Beobachtung sein. Vielmehr sind geringe Prüfungsumfänge vor allem zu beobachten, wenn bereits mit wenigen Elementen eine wertmäßige Abdeckung des Prüffelds von mehr als 50% erreicht wurde. Die mengenmäßige Abdeckung der Prüffelder ist aufgrund der bei der bewussten Auswahl sowie beim MUS vorherrschenden Fokussierung auf höherwertige Elemente erwartungsgemäß deutlich geringer.

Mit einer Spannweite von bis zu 75 Elementen (bewusste Auswahl), 125 (statistische Heuristik) sowie 166 (MUS) liegen die Prüfungsumfänge im Mittel über den z.B. gemäß AICPA Audit Guide geforderten Umfängen.[127] Zudem liegen die mittels Heuristik ermittelten Umfänge in etwa auf dem gemäß *IDW PS 310*, Tz. A10, geforderten Niveau.

Ein deutlich positiver Zusammenhang lässt sich zwischen der Höhe des inhärenten sowie des Kontrollrisikos und dem Prüfungsumfang feststellen. Parallel dazu steigen die Wahrscheinlichkeit des Einsatzes repräsentativer Stichproben und die Höhe der erreichten Abdeckungsraten in Prüffeldern mit einem hohen Kontrollrisiko. Dieses Ergebnis ist vor dem Hintergrund des in risikoträchtigen Prüffeldern nur in geringem Maße vorhandenen Vorwissens zur Fehlerhypothese zu begrüßen.

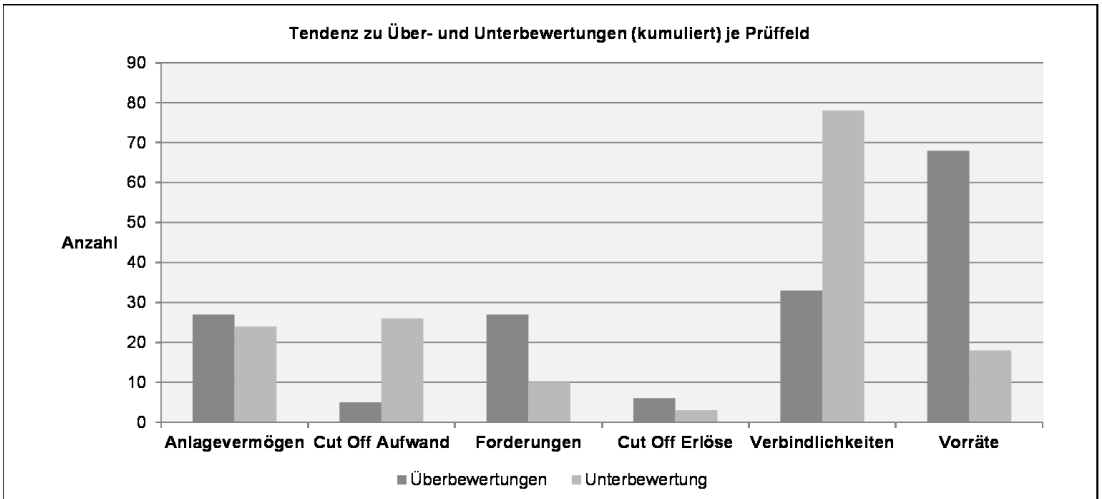
[127] Vgl. AICPA (Hrsg.), Audit Guide – Audit Sampling, New York 2014, F Chapter 7; Appendix C.

Tabelle 4.3: Prüfungsumfänge und erreichte Abdeckung

			Anlagevermögen	Cut Off Aufwand	Forderungen	Cut Off Erlöse	Verbindlichkeiten	Vorräte
P (wertmäßiger Umfang des Prüffelds in T€ (Grundgesamtheit))	**Bewusst**	**Mittelwert**	3.260	34.171	2.779	50.995	2.751	9.423
		Median	743	2.065	1.114	3.877	739	2.630
	Repräsentativ	**Mittelwert**	4.166	9.939	14.302	6.159	66.640	8.700
		Median	3.327	3.563	2.304	3.131	1.650	3.436
N (Zahl der Elemente in der Grundgesamtheit)	**Bewusst**	**Mittelwert**	392	2.925	271	3.073	736	2.225
		Median	99	844	105	466	313	291
	Repräsentativ	**Mittelwert**	962	15.340	1.455	4.992	1.150	4.531
		Median	490	2.384	791	1.406	601	2.073
p (wertmäßiger Umfang ausgewählter Elemente in T€)	**Bewusst**	**Mittelwert**	1.343	12.420	1.412	11.842	1.404	869
		Median	384	344	485	597	302	430
	Repräsentativ	**Mittelwert**	1.720	1.116	7.147	1.564	81.333	1.330
		Median	572	525	674	547	684	392
N (Zahl der ausgewählten Elemente)	**Bewusst**	**Minimum**	1	2	1	2	2	1
		Maximum	50	41	21	75	40	60
		Mittelwert	10	21	11	17	12	16
		Median	7	14	5	11	9	10
	Repräsentativ	**Minimum**	2	10	5	2	4	6
		Maximum	23	82	166	125	45	83
		Mittelwert	16	32	25	27	15	35
		Median	9	21	16	20	14	24
Abdeckung (Mittelwert)	**Bewusst**	**Wert**	59%	42%	64%	38%	69%	36%
		Menge	16%	3%	16%	11%	4%	12%
	Repräsentativ	**Wert**	37%	21%	47%	23%	48%	20%
		Menge	9%	5%	10%	4%	6%	3%

4.6.2.3 Fehlercharakteristika

Grundgesamtheiten im Rechnungswesen werden regelmäßig sehr geringe Fehlerraten zugesprochen.[128] Entgegen dieser Feststellung im anglo-amerikanischen Raum zeigt die Studie jedoch, dass die Aufdeckung wesentlicher Fehler im Rahmen von Auswahlprüfungen durchaus üblich ist (Abbildung 4.5). So ergibt sich in 82 von 340 Fällen die Aufdeckung mindestens einer falschen Angabe und global betrachtet eine mengenmäßige Fehlerhäufigkeit (Prüffeld enthält mindestens einen Fehler) von 24,1%. Werden allein repräsentative Stichproben betrachtet, liegt diese Kennzahl bei 7,8%. Die bewusste Auswahl liefert damit erwartungsgemäß eine höhere Aufdeckungsrate.



Fehlereigenschaften je Prüffeld

		Anlagevermögen	Vorräte	Forderungen	Verbindlichkeiten	Cut Off Erlöse	Cut Off Aufwand
Anzahl Fehler	Minimum	0	0	0	0	0	0
	Maximum	8	17	6	16	2	7
	Mittelwert	2	4	2	4	1	2
	Median	1	2	2	2	1	1
	Tendenz	ausgeglichen	überbewertet	überbewertet	unterbewertet	überbewertet	unterbewertet
Fehlerhöhe kumuliert in T€	Mittelwert	5,1	28,9	128,2	2,4	23,0	-16,1
	Median	0,2	5,5	1,0	-0,2	21,0	-2,0

Abbildung 4.5: Eigenschaften entdeckter Fehler

Die Untersuchung zeigt hinsichtlich der Fehlerraten, dass diese bei allen Prüffeldern regelmäßig unter 1% liegen und somit als gering einzustufen sind. Darüber hinaus lassen sich folgende Feststellungen zu aufgedeckten Fehlern treffen (Abbildung 4.5):

[128] Dieser Effekt hat laut empirischen Untersuchungen in den USA in der post-SOX-Periode weiter zugenommen; vgl. Durney/Elder/Glover, Auditing 2014, S. 80 f.; deskriptiv auch Ruhnke/von Torklus, WPg 2008, S. 1126.

Die Posten Vorräte und Verbindlichkeiten weisen durchschnittlich höhere Fehlerraten als die übrigen Posten auf.

- Wertmäßig liegen die Fehlerraten auf Seiten der Forderungen/Erlöse und Vorräte höher als bei den verbleibenden Prüffeldern.
- Bei wesentlichen Fehlern im Bereich der Forderungen/Erlöse und Vorräte handelt es sich vorwiegend um Überbewertungen, während Verbindlichkeiten/Aufwendungen tendenziell unterbewertet sind.

4.6.2.4 Handhabung nicht fehlerfreier Prüffelder

IDW PS 300 n.F. und *IDW PS 310* zeigen im Falle entdeckter Fehler Konsequenzen bezüglich der Ausweitung des Prüfungsumfangs, der Fehlerhochrechnung, der Ziehung von Ersatzelementen sowie der möglichen Fehlerisolierung auf. Die Untersuchung ermöglicht dazu diverse Feststellungen (Tabelle 4.4).

Tabelle 4.4: Behandlung nicht fehlerfreier Prüffelder

	Zahl der Prüfungen mit Fehlern (relativ)	Hochrechnungen	Fehler-isolierungen	Revisionen
Anlagevermögen	10	6	1	–
Vorräte	21	13	9	8
Forderungen	10	10	2	–
Verbindlichkeiten	21	17	2	–
Cut Off Erlöse	8	1	4	4
Cut Off Aufwand	12	4	3	5

Bei Verwendung repräsentativer Auswahlverfahren ist gemäß *IDW PS 310* stets eine Berücksichtigung des Stichprobenrisikos bzw. eine Hochrechnung notwendig. In 76,2% der betrachteten Fälle hat eine solche Berücksichtigung stattgefunden. Die Fälle, in denen die Hochrechnung unterblieb, betreffen vor allem die Prüfung der Periodenabgrenzung sowie der Vorräte. Generell sollte immer – auch bei der bewussten Auswahl – eine Aussage bezüglich des erwarteten absoluten Fehlers im Prüffeld in Relation zur postenbezogenen Wesentlichkeit getroffen werden, da allein dies eine Aussage über die Ordnungsmäßigkeit des gesamten Prüffelds erlaubt.

Die in der Fachliteratur angeführte Kritik,[129] dass auch bei Aufdeckung wesentlicher Fehler häufig keine Revision des Prüfungsumfangs stattfindet, kann nicht bestätigt werden. Infolge der Aufdeckung falscher Angaben hat in ca. 84% der Fälle eine Ausweitung der

[129] Vgl. Mauldin/Wolfe, Contemporary Accounting Research 2014, S. 663 f.

Prüfungshandlungen stattgefunden. Die Ziehung von Ersatzelementen aufgrund der Unmöglichkeit einer Einholung des Prüfungsnachweises konnte nicht beobachtet werden. In den seltenen Fällen der Nicht-Anwendbarkeit der Prüfungshandlung auf das Ursprungselement wurde stets ein Ersatzelement ausgewählt.

Insgesamt entspricht der Umgang mit nicht fehlerfreien Prüffeldern damit fast ausnahmslos den Anforderungen von *IDW PS 300 n.F.* und *IDW PS 310*.

4.6.3 Kritische Würdigung

Zu diskutieren ist vorrangig der Ansatz selektiver Auswahlmethoden. Werden keine Fehler gefunden, kann dies bei einer bewussten Auswahl der Prüfelemente aus zweierlei Gründen resultieren:

- entweder aufgrund weniger Fehler in der Grundgesamtheit
- oder aufgrund einer falschen Fehlerhypothese bzw. einer nicht korrekten Verwertung von Vorwissen, was zu einer verzerrten bewussten Auswahl führt.

Beide Fälle lassen sich weder nachträglich vom Prüfer noch durch eine etwaige Qualitätskontrolle nachvollziehen. Generell darf aufgrund der Kompetenz des Abschlussprüfers jedoch davon ausgegangen werden, dass unter Effizienzgesichtspunkten die bewusste Auswahl eher als eine reine Zufallsauswahl in der Lage ist, (wesentliche) Fehler zu identifizieren.

4.7 Implikationen für die Praxis

Stichprobenverfahren werden aufgrund ihrer mittels Software vermeintlich einfachen Anwendung schnell zur „Black Box“. *IDW PS 310* sensibilisiert daher stärker als die *Stellungnahme HFA 1/1988* für potenzielle Schwachstellen bei der Durchführung von Stichprobenverfahren. Die Studie zeigt, dass in diesen Bereichen durchaus Verbesserungspotenzial festzustellen ist. Praktisch ergibt sich vor allem bei hochgerechneten Fehlern das Problem, geschätzte Korrekturbedarfe einzelnen Posten des Prüffelds zuzuordnen.

Auswahlverfahren müssen stets auf die den Prüffeldern immanenten Eigenschaften zugeschnitten sein. Stichproben im Sinne von *IDW PS 310* bieten sich an, wenn Geschäftsvorfälle ohne geeignete Vorinformationen vorliegen, was jedoch in der Praxis – fernab von Massengeschäften und Umsatzerlösen, die stets bedeutsame Risiken darstellen –

selten der Fall sein wird.[130] Sie bieten sich darüber hinaus an, wenn das interne Kontrollsystem effektiv ist, da der notwendige Stichprobenumfang dann tendenziell geringer ausfallen kann.

Besteht ein erhöhtes Kontrollrisiko, wird die gezielte Auswahl risikobehafteter Elemente zweckmäßig sein. Gemäß *IDW PS 300 n.F.* handelt es sich bei der bewussten Auswahl von Prüfobjekten nicht (mehr) um eine Stichprobe. Dies ist bei der Prüfungsplanung und -durchführung sowie der Dokumentation zu würdigen. In Zukunft wird vermehrt auf die Verwendung mathematisch-statistischer Termini in den Arbeitspapieren zu achten sein, sofern Stichprobenverfahren im Sinne von *IDW PS 310* Anwendung finden.

4.8 Zusammenfassung

Statistische und nicht-statistische Stichproben gehören zum Handwerkszeug nahezu jeder Abschlussprüfung. Es ist vor dem Hintergrund der technischen und methodischen Entwicklung zu begrüßen, dass mit *IDW PS 300 n.F.* und *IDW PS 310* eine Aktualisierung der normativen Vorgaben zur Verwendung von Auswahlprüfungen erfolgt.

Der Kern etwaiger Probleme der bewussten und nicht-statistischen Auswahlmethoden liegt weniger in der prozessorientierten Umsetzbarkeit vorgesehener Arbeitsabläufe. Vielmehr ist es die gemäß den Prüfungsstandards geforderte kognitive Fähigkeit des Anwenders, statistische Gesetzmäßigkeiten in den intuitiven Ansatz zu übernehmen. Die Menge der sich in diesem Zusammenhang überlagernden Denkprozesse[131] ist hoch und anfällig für eine mangelhafte Anwendung der neuen Prüfungsstandards. Die Entwicklung unterstützender Werkzeuge (decision aids) – wie sie im amerikanischen Berufsstand bereits Anwendung finden –[132] scheint sinnvoll.[133]

Die im Vergleich zur *Stellungnahme HFA 1/1988* pragmatischere und stärker praxisorientierte Herangehensweise dürfte zu einer größeren Sensibilisierung der Anwender für die Schwachstellen von Auswahlprüfungen beitragen und einen tendenziell positiven Einfluss auf die Prüfungsqualität haben.

[130] Vgl. Hömberg, BFuP 1997, S. 248 f.
[131] Vgl. ausführlich Mochty, in: Richter (Hrsg.), Entwicklungen der Wirtschaftsprüfung, Bielefeld 2003, S. 188.
[132] Vgl. Bonner/Libby/Nelson, The Accounting Review 1996, S. 221 (im Allgemeinen), sowie Messier/Kachelmeier/Jensen, Auditing 2001, S. 81–96 (im Besonderen). Letztere zeigen, dass Leitfäden zur Unterstützung von Auswahlprüfungen zu einer höheren Prüfungsqualität führen.
[133] So auch Heese/Braatsch, WPg 2013, S. 848.

5 Prüferisches Ermessen vs. mathematische Präzision – Schränkt IDW PS 310 die Handlungsfreiheit des Abschlussprüfers ein?

5.1 Publikationsdetails

Zusammenfassung: IDW PS 300 n. F. sowie IDW PS 310 wurden vom Hauptfachausschuss des IDW am 29. Juli 2016 verabschiedet, womit die Durchführung von Auswahlprüfungen – insbesondere Stichproben – für Abschlussprüfungen von Geschäftsjahren, die nach dem 15.12.2016 beginnen, neu reguliert wird. Neben der Ablösung von IDW PS 300 a. F. zur Erlangung von Prüfungsnachweisen sowie HFA 1/1988 halten hierdurch die Inhalte der ISA 500 und ISA 530 Einzug in sämtliche handelsrechtliche Jahresabschlussprüfungen.

Koautoren: Christoph Oldewurtel.

Stichwörter: IKS, IDW PS 310, Audit Sampling, Decision Aids, Repräsentative Auswahl.

Publikationsstatus: Veröffentlicht in: *WP Praxis* 6 (7): 144–149. Ebenfalls erschienen in WP-Praxis 2018, Sonderausgabe "Inventurprüfung und Stichprobenverfahren - Werkzeuge für die Jahresabschlussprüfung": 15–20.

5.2 Einleitung

Die Erlangung ausreichender und angemessener Prüfungsnachweise ermöglicht es dem Abschlussprüfer ein Urteil über die Ordnungsmäßigkeit des Jahresabschlusses sowie die zugrundeliegende Buchführung seines Mandanten zu fällen. Solange der Jahresabschluss nicht durch die ausschließliche Anwendung systemgestützter und analytischer Prüfungshandlungen weitestgehend automatisiert geprüft werden kann, stellt die Durchführung von Einzelfallprüfungshandlungen einen unabdingbaren Eckpfeiler eines jeden Prüfungsprozesses dar. Die Menge potenziell prüfbarer Geschäftsvorfälle und damit der Umfang auswertbarer Daten steigen dabei nicht nur stetig an, sondern sind aufgrund komplexerer IT- und Prozessstrukturen innerhalb der zu prüfenden Unternehmen auch zunehmend als abstrakt zu klassifizieren. Ungeachtet jener Erschwernisse obliegt es dem Abschlussprüfer, eine angemessene Auswahl der Prüfelemente zu tätigen, als auch die Integrität – insbesondere Richtigkeit und Vollständigkeit – der betrachteten Daten sicherzustellen. Die im vergangenen Jahr verabschiedeten Prüfungsstandards IDW PS 300 n. F. und IDW PS 310[134] konkretisieren jene Erlangung von Prüfungsnachweisen sowie die Nutzung repräsentativer Auswahlverfahren (Stichproben) im Sinne ihrer internationalen Gegenstücke ISA 500 sowie ISA 530.[135] IDW PS 310 ersetzt zudem die für Stichproben bisher gültige Stellungnahme HFA 1/1988.[136] Der vorliegende Beitrag dient der Ergänzung bisheriger Publikationen zur Durchführung von Auswahlprüfungen und beleuchtet insbesondere die Konkretisierung der Inhalte durch die vom IDW veröffentlichten Fragen & Antworten zu den beiden Prüfungsstandards. Ergänzend wird auf weitere Quellen zur Unterstützung des Prüfers bei der Durchführung von Stichprobenprüfungen hingewiesen, um auf diesem Wege potenzielle Anlaufstellen im Falle sich ergebender etwaiger Anwendungsprobleme in praxi aufzuzeigen.

5.3 Auswahlprüfungen in der Prüfungsliteratur

IFAC und IDW verzahnen das Prüfungsrisiko eng mit dem Prüfungsumfang und widmen dem Themenkomplex der Stichprobenverfahren jeweils dezidierte Prüfungsstandards.

[134] IDW PS 300 n. F. „Prüfungsnachweise im Rahmen der Abschlussprüfung“, IDW Life 2016, S. 624 ff.; IDW PS 310 „Repräsentative Auswahlverfahren (Stichproben) in der Abschlussprüfung“, IDW Life 2016, S. 636 ff.

[135] ISA 500 „Audit Evidence“ und ISA 530: „Audit Sampling”, in: IFAC (Hrsg.), 2015 Handbook of International Quality Control, Auditing, Review, Other Assurance and Related Services Pronouncements, New York 2015, Volume I.

[136] IDW, HFA Stellungnahme 1/1988: Zur Anwendung stichprobengestützter Prüfungsmethoden bei der Jahresabschlussprüfung, WPg 1988, S. 240 ff.

Konkret findet die Darstellung verfügbarer Auswahlmethoden für Prüfungselemente seit Beendigung des Clarity Project der IFAC in ISA 500 Berücksichtigung, wohingegen Stichprobenverfahren gesonderte Behandlung in ISA 530 erfahren. Die Inhalte von ISA 530 sind dabei in erheblichem Umfang durch die Verlautbarungen der AICPA, insbesondere SAS No. 39 respektive der aktuellen AU Section 530 nebst deren Interpretation in AU Section 9350 motiviert, ohne jedoch die gleiche Regelungstiefe anzustreben. Der gleichen Struktur wie die ISA folgen auch die nunmehr reformierten deutschen Prüfungsstandards (vgl. Abbildung 5.1). Danach verweist IDW PS 300 n. F. im Rahmen der Darstellung verfügbarer Auswahlmethoden zur Einholung von Prüfungsnachweisen in Stichproben auf deren ausschließliche Behandlung in IDW PS 310.[137] Im Mittelpunkt sämtlicher Verlautbarungen steht dabei die Möglichkeit der Schlussfolgerung bzw. Inferenzbedingung von Stichprobenverfahren. Letztere werden danach primär durch die mögliche Hochrechnung entdeckter falscher Angaben bzw. Kontrollabweichungen[138] auf die Grundgesamtheit definiert.

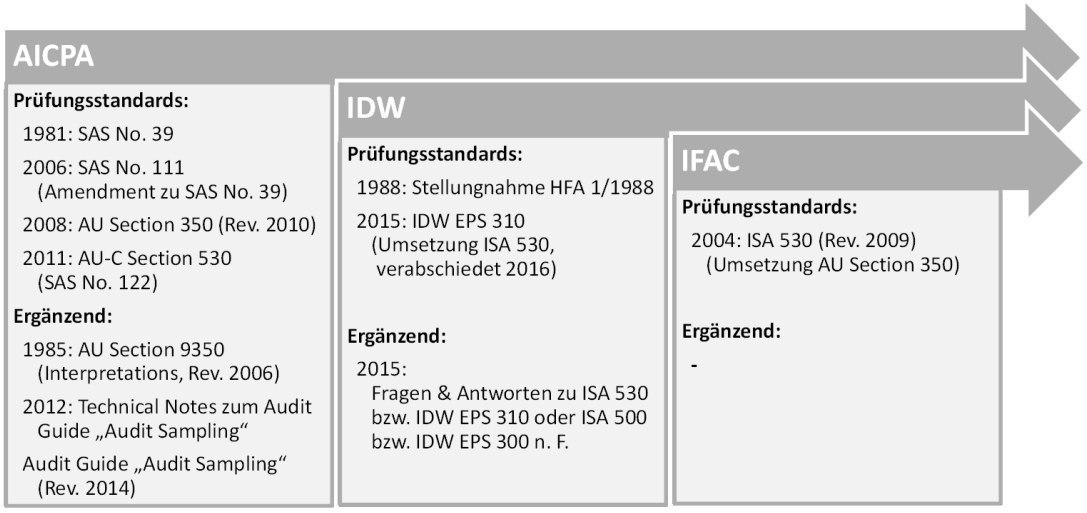


Abbildung 5.1: Verlautbarungen von AICPA, IFAC und IDW zu repräsentativen Auswahlverfahren

5.3.1 Diskurs im Schrifttum

Der Einsatz von Auswahlprüfungen stellt seit jeher einen „ungelösten Dauerbrenner“[139] in den Reihen des nationalen wie auch internationalen Schrifttums dar.[140] Dabei konnten

[137] Vgl. IDW PS 300 n. F., Tz. A52.
[138] Im Folgenden wird der Begriff „Fehler“ als Synonym für Kontrollabweichungen und falsche Angaben verwendet.
[139] Mochty, in: Richter (Hrsg.), Entwicklungen der Wirtschaftsprüfung: Prüfungsmethoden - Risiko - Vertrauen, Bielefeld 2003, S. 185.
[140] Vgl. Elder et al., Auditing 2013, Suppl. 1, S. 99-129.

bisherige Publikationen zur Thematik zeigen, dass die Einhaltung der Inhalte der ISA 500 und ISA 530 sowie die schlussendliche Einbettung in Prüfungsansätze im Rahmen der praktischen Anwendung keineswegs einheitlich verlaufen und in bestimmten Fällen zu mangelnder Prüfungsqualität führen können.[141] Hinsichtlich verwendbarer und normativ legitimierter Methoden lässt sich auch im deutschen Raum kein einheitlicher Konsens finden, wobei normativ zudem bis zum Jahr 2015 eine gewisse Stagnation zu konstatieren war, während von Seiten der Wissenschaft fortwährend neue vermeintlich effiziente Auswahlverfahren entwickelt wurden, deren Übernahme in die Berufspraxis jedoch bislang nahezu ausblieb.[142]

Es mag daher nicht verwundern, dass nach der Veröffentlichung von IDW EPS 300 n. F. und IDW EPS 310 im Mai 2015 eine Vielzahl an (teilweise kritischen) Stellungnahmen[143] aus Wirtschaft, Wissenschaft und Prüfungspraxis den Diskurs über die überarbeiteten bzw. neu gefassten Standards beherrschte. Es erwächst folglich die Frage, worauf bei einer Erstanwendung der neuen Prüfungsstandards im Wesentlichen zu achten ist und welche Regelungsinhalte den Prüfer ohne weitere Konkretisierung vor potenzielle Hürden stellen.

5.3.2 Aktueller Stand der Normen für handelsrechtliche Jahresabschlussprüfungen

Im Fokus jener Stellungnahmen zu IDW EPS 300 n. F. und IDW EPS 310 stand dabei unter anderem die auch innerhalb der endgültigen Standards analog zu ISA 530 vertretene Regelung, dass im Falle einer bewussten Auswahl zu prüfender Elemente eines Prüffelds künftig nicht mehr von dem Begriff „Stichprobe“ Gebrauch gemacht werden darf. Das IDW reagierte auf die Fülle an Stellungnahmen vor allem mit der Veröffentlichung einer umfangreichen Sammlung von „Fragen & Antworten“ (im Folgenden auch: „F&A“), welche Missverständnissen vorbeugen und den Abschlussprüfer auf die nahende Erstanwendung vorbereiten soll.[144] Diese zeitnah nach Veröffentlichung der Entwürfe der neuen

[141] Vgl. Christensen/Elder/Glover, Accounting Horizons 2015, S. 61-82.
[142] Vgl. Baumeister/Oldewurtel, WP Praxis 2016, S. 169-175; Baumeister/Oldewurtel, WP Praxis 2016, S. 193-198.
[143] Vgl. u.a. AK Rechnungslegung des Deutschen Steuerberaterverbands e.V., Stellungnahme B 11/15 (https://www.idw.de/blob/86276/dc89972e5189794987616f7baa9b1797/down-idweps310-dstv-data.pdf) (Abruf am 03.04.2017).
[144] Vgl. Fragen und Antworten: Zur Durchführung einer repräsentativen Auswahl (Stichproben) nach ISA 530 bzw. IDW EPS 310 oder einer bewussten Auswahl nach ISA 500 bzw. IDW EPS 300 n.F., IDW Life 2016, S. 91 ff.

Prüfungsstandards publizierten Inhalte ermöglichen folglich zumindest potenziell eine frühzeitige Sensibilisierung der schlussendlichen Anwender mit den wesentlichen Neuerungen von IDW PS 300 n. F. und IDW PS 310. So finden sich in den F&A neben ausführlichen Definitionen und Begriffsabgrenzungen auch konkrete Anwendungshinweise mit Beispielen, wobei sich der Großteil der Erläuterungen auf die durch IDW PS 310 regulierten repräsentativen Auswahlverfahren bezieht. Als Beitrag zur Vermeidung etwaiger Unstimmigkeiten sowie als Hinweise zur praktischen Durchführung von Auswahlprüfungen kommen jene F&A insbesondere kleineren Prüfungsgesellschaften zugute.[145]

5.4 Hilfreiche Quellen zur Durchführung von Auswahlprüfungen in der Jahresabschlussprüfung

Grundlagen- und Spezialliteratur der allgemeinen statistischen Methodenlehre eignet sich nur bedingt als Informationsquelle zur Erhebung und Auswertung von Stichproben im Anwendungsfall der Jahresabschlussprüfung, da Prüffelder im Rechnungswesen deutlich andere statistische Merkmale aufweisen als beispielsweise Grundgesamtheiten in den Bereichen der Marktforschung oder der Naturwissenschaften. So ist insbesondere die Tatsache, dass die Ausprägungen des interessierenden Untersuchungsmerkmals (die Menge an Kontrollabweichungen respektive die Häufigkeit und Höhe falscher Darstellungen) verhältnismäßig selten in der Grundgesamtheit vorkommen, als charakteristisches Merkmal im Rahmen der Jahresabschlussprüfung zu untersuchender Datenmengen anzuführen. Fehlerraten liegen häufig bei weit unter einem Prozent und folgen i. d. R. keiner Normalverteilung.[146] Aufgrund dieser Eigenschaften stoßen klassische Stichprobenverfahren in der Abschlussprüfung häufig an ihre Grenzen. Es erscheint daher ratsam, zur Prüfungsvorbereitung und -durchführung weitere ergänzende Literatur heranzuziehen. Neben den bereits benannten F&A des IDW stellt auch der amerikanische Standardsetter in Form des AICPA Audit Guide „Audit Sampling“ zwei konzeptionell sehr unterschiedliche Hilfestellungen bereit, die im Folgenden skizziert werden.

5.4.1 Verlautbarungen des IDW zu PS 300 n. F. und PS 310

Der Komplexitätsgrad einer Auswahlprüfung hängt, neben weiteren Variablen, in hohem Maße vom Datentyp des interessierenden Untersuchungsmerkmals ab, wobei sich

[145] Vgl. Christensen/Elder/Glover, a.a.O. 2015, S. 61-81, wonach Big 4 und second tier-Gesellschaften eigene Ansätze zum Umgang mit Auswahlverfahren entwickeln.
[146] Vgl. Ham/Losell/Smieliauskas, CAR 1987, S. 215.

grundsätzlich zweierlei Fälle unterscheiden lassen: Einerseits die Attribut-Stichprobe zur Prüfung bei binärer Ausprägung des Untersuchungsmerkmals, welche regelmäßig bei Funktionsprüfungen des internen Kontrollsystems Anwendung findet. Andererseits die Variablen-Stichprobe bei kardinaler Skalierung der Ausprägung des Untersuchungsmerkmals, welche regelmäßig bei aussagebezogenen Einzelfallprüfungen in monetären Prüffeldern Relevanz entfaltet.

IDW PS 310 folgt dieser Aufteilung und stellt einige Inhalte (z. B. die Möglichkeit der Hochrechnung des ermittelten Abweichungsgrades auf die Grundgesamtheit[147] oder die Wahl eines angemessenen Stichprobenumfangs[148]) daher getrennt dar. Praktisch gestaltet sich die Durchführung einer Variablen-Stichprobe ungleich komplexer als die einer Attribut-Stichprobe. Letztere erfolgt regelmäßig auf Basis der einfachen Zufallsauswahl mit fixen Stichprobenumfängen in Abhängigkeit der Kontrollfrequenz. Der ermittelte Kontrollabweichungsgrad der Stichprobe entspricht dabei dem besten Schätzer für den wahren Kontrollabweichungsgrad der Grundgesamtheit. Für die Variablen-Stichprobe gilt hingegen, dass einzelne Prüfelemente mit einem Abweichungsgrad kleiner 100 % vorliegen können. Es existiert aus diesem Grund eine Vielzahl möglicher Erhebungs- und Auswertungsverfahren.

Die F&A widmen sich daher vor allem der Durchführung von Variablen-Stichproben. Bezüglich der bewussten Auswahl wird zuvorderst, wie bereits geschildert, klargestellt, dass sie anders als Stichproben unter keinen Umständen repräsentativ für die Grundgesamtheit sein kann, gleichwohl jedoch häufig eine adäquate Alternative darstellt, da sie in hohem Maße das Vorwissen des Prüfers zu nutzen vermag (für eine Zusammenfassung der Vor- und Nachteile der unterschiedlichen Auswahlverfahren vgl. Tabelle 5.1).[149] Aus letzterer Feststellung wird abgeleitet, dass es für eine bewusste Auswahl keine quantitativen Vorgaben zum Mindestprüfungsumfang geben kann, da dieser von qualitativ zu würdigenden Parametern abhängt.[150]

[147] Vgl. IDW PS 310, Tz. A18.
[148] Vgl. IDW PS 310, Anlagen 2 und 3.
[149] F&A, 3.1, 2.2 i. V. m. 3.2 und 3.5; die bewusste Auswahl ist ebenfalls kein nichtstatistisches Verfahren, vgl. F&A, 4.5.
[150] Vgl. für Variablen-Stichproben F&A, 3.3., zu Funktionsprüfungen des IKS auch F&A, 9.3; zu Parametern auch IDW PS 310, Anlagen 2 und 3.

Tabelle 5.1: Vor- und Nachteile der bewussten und repräsentativen Auswahl

Repräsentative Auswahl			Bewusste Auswahl
+	Ermittlung eines effizienten Stichproben-umfangs (wenig over-/underauditing)	+	Einfluss von Erfahrungswerten und Vorwissen des Prüfers
	Quantitative Darstellung des Prüfungsergebnisses und höhere Wahrscheinlichkeit der Identifikation eines (zu) hohen Stichprobenrisikos		Anpassung auf individuelle Bedürfnisse
	Messbarkeit des Stichprobenrisikos (gesteigertes Vertrauen in Prüfungsergebnis)		Bei vorhandenen Vorkenntnissen effektiv und effizient
	Messbarkeit der Güte von Prüfungsnachweisen		Kein Training notwendig
		-	Kein Repräsentationsschluss quantifizierbar
	Festlegung von Sicherheitsgrad und Präzision		Keine objektive Messbarkeit des Stichprobenrisikos
	(Theoretisch) Einfache Exkulpation durch Objektivierbarkeit des Ergebnisses		Risiko eines zu hohen/zu geringen Prüfungsumfangs
	Prüfungsprozess und Dokumentation konsistenter und einfacher zu standardisieren		Qualität abhängig von subjektiver Urteilsfähigkeit des Prüfers
-	Kann zeit- und kostenintensiv sein		Gezieltes Training nur eingeschränkt möglich
	Schulungskosten für Mitarbeiter		
	Anschaffungskosten für Software		

Zu Konzeption und Umfang von Stichproben wird dargelegt, worin der Unterschied zwischen statistischen und nichtstatistischen Verfahren besteht: Insbesondere die in der Praxis häufige Verwendung einer lediglich zufallsimitierenden Auswahl sowie der Einsatz von Hilfstabellen zur Ermittlung eines näherungsweisen mathematisch-statistischen Stichprobenumfanges führen dazu, dass nichtstatistische Verfahren dominieren. Auf dieser Basis kann entsprechend keine quantitative Berücksichtigung des Prüfungsrisikos oder eine statistische Hochrechnung auf die Grundgesamtheit erfolgen.[151] In den F&A wird demgegenüber jedoch ebenfalls klargestellt, dass bei Nichtvorliegen einer vollständigen und entsprechend mathematisch auswertbaren elektronischen Datenbasis häufig nichtstatistischen Verfahren der Vorrang gewährt werden kann.[152]

[151] Vgl. F&A, 4.1 zur Variablen-Stichprobe; für Funktionsprüfungen des IKS wird die zufallsimitierende Auswahl als Mindestanspruch angesehen.
[152] Vgl. F&A, 4.3.

Auch im Falle einer repräsentativen Auswahl sprechen sich die F&A nicht für eine quantitative Empfehlung von Mindestumfängen aus, wobei jedoch Näherungswerte für Funktionsprüfungen des IKS genannt werden.[153] Für den Praktiker werden die zum Teil recht sperrigen Inhalte des IDW PS 310 etwas anschaulicher zusammengefasst, insbesondere hinsichtlich der Berücksichtigung des Wesentlichkeitsgrundsatzes sowie der Fehlererwartung an die Grundgesamtheit.[154]

Hinsichtlich verwendbarer Auswahlmethoden wird hingegen explizit dargelegt, dass die praktisch häufig vorkommende zufallsimitierende Auswahl (bisher „Auswahl aufs Geratewohl", gemäß F&A, 4.3 z. B. bei der Prüfung der Stichtagsinventur) i. S. des IDW PS 310, Anlage 4 keine legitime Auswahlmethode für statistische Stichproben darstellt.[155] Letztere verlangen nach unabhängigen Auswahlverfahren und somit nach entweder einer echten oder systematischen Zufalls- oder nach einer wertproportionalen Auswahl (Monetary Unit Sampling).[156]

Aufgrund ihrer hohen praktischen Relevanz erfährt die wertproportionale Auswahl eine ausführliche Würdigung in den F&A, wobei neben möglichen Einsatzbereichen auch die Limitationen entsprechend erläutert werden. Hierzu zählt insbesondere, dass eine Nutzung für Passivposten bzw. bei Vermutung auf Unterbewertung oder Unvollständigkeit unterbleiben sollte.[157]

Schlussendlich wird neben der Darstellung der verfügbaren Verfahren auch festgehalten, wie im Falle des Auffindens von Kontrollabweichungen bzw. falschen Darstellungen im Rahmen von Stichproben vorzugehen ist[158] und wie verfahren werden sollte, sofern einzelne Elemente einer Auswahl nicht prüfbar sind (bei Unmöglichkeit oder Nichtanwendbarkeit der Prüfungshandlung) oder als sogenannte „Anomalie" klassifiziert und damit vom erwarteten Fehler in der Grundgesamtheit isoliert werden. An eine Fehlerisolierung knüpfen sich dabei strenge Anforderungen sowie ggf. die Prüfung auf das Vorliegen fraudulenter Handlungen.[159]

[153] Vgl. für monetäre Prüffelder F&A, 4.8, für Funktionsprüfungen des IKS F&A, 9.2, 9.3.
[154] Vgl. IDW PS 310, Anlagen 2 und 3.
[155] Vgl. begründend dazu auch F&A, 5.7 sowie Hall/Hunton/Pierce, Behavioral Research in Accounting 2000, S. 231-255.
[156] Vgl. F&A, 5.1.
[157] Vgl. F&A, 5.3, 5.4.
[158] Vgl. zur Hochrechnung von falschen Darstellungen bzw. zum Kontrollabweichungsgrad F&A, 7.1 ff. bzw. 9.5. Eine Hochrechnung kann bei nichtstatistischen Methoden gem. F&A, 8.3 auch qualitativ im Rahmen des prüferischen Ermessens erfolgen.
[159] Vgl. F&A, 6.5.

Die F&A geben in Abschnitt 8 zudem Hinweise, wie Fehler in einer Stichprobenprüfung, die hochgerechnet oberhalb einer tolerierbaren Grenze liegen, zu würdigen sind. Im Falle von Funktionsprüfungen führt dies zwangsläufig zur Notwendigkeit der Revision der Risikoeinschätzung des Prüfers.[160] Im Falle von falschen Darstellungen in monetären Prüffeldern kann die Schlussfolgerung einer überschrittenen Fehlererwartung hingegen dazu führen, dass die Stichprobe ausgeweitet werden sollte, der Mandant die Fehler korrigieren muss oder das Prüffeld als nicht ordnungsmäßig abgelehnt werden sollte.[161]

Insgesamt verfolgen die F&A einen äußerst abstrakten Weg, die Inhalte aus IDW PS 300 n. F. und PS 310 zu konkretisieren. Verbale Erläuterungen überwiegen, eine quantitative Hilfestellung wird nahezu nicht geleistet. Inwieweit hiervon folglich der Anwenderkreis insbesondere in Form kleinerer Prüfungsgesellschaften und -Sozietäten zu profitieren weiß, entzieht sich in Ermangelung empirischer Studien einem fundierten Urteil. Obschon das IDW vorsieht, die F&A bei Bedarf zu aktualisieren, scheint dies unwahrscheinlich, da auch die finalisierten Standards zum benannten Themenkomplex nahezu exakt den Entwurfsstandards entsprechen.

5.4.2 AICPA Audit Guide "Audit Sampling"

Neben den Verlautbarungen des nationalen wie auch internationalen Standardsetters kommen aufgrund der Anlehnung der Inhalte von ISA 530 bzw. IDW PS 310 an AU Section 530 zugleich die Publikationen des AICPA als potenziell nutzenstiftende weiterführende Literatur auch für Anwender der deutschen Prüfungsnormen in Betracht. Dies gilt umso mehr aufgrund der Tatsache, dass die F&A, wie dargelegt, keine Decision Aids und nur eingeschränkt quantitative Hilfestellung leisten. Aus Sicht des Praktikers, der sich statistischer und nichtstatistischer Stichproben bedienen möchte, kommt in diesem Zusammenhang insbesondere der Audit Guide „Audit Sampling" in Betracht, weicht selbiger in Ausrichtung und Regelungstiefe doch von den F&A ab. So werden auf nahezu 200 Seiten zwar ähnliche Inhalte wie die des IDW PS 310 behandelt; dies geschieht jedoch in erheblicher Breite: Beispielsweise wird im Audit Guide u. a. eine umfangreiche Negativabgrenzung allein solcher Prüfungshandlungen vorgenommen, die keine Stichproben darstellen und es werden entsprechende korrespondierende Beispiele angeführt.[162]

[160] Vgl. F&A, 8.2.
[161] Vgl. F&A, 8.5 sowie zur weiteren Konsequenz hinsichtlich der Aufzeichnung, Korrektur und Kommunikation mit dem Mandanten F&A, 8.6 sowie IDW PS 250 n. F.
[162] Vgl. AICPA, a.a.O. 2014, Kap. 1, Rn. 6-20.

Während IDW PS 310 sowie die F&A gleichermaßen Definitionen festlegen und die grundsätzliche Vorgehensweise bei der Stichprobenerhebung aufzeigen, geht der Audit Guide über diese Inhalte hinaus: Er beinhaltet eine vollständige Prozessbeschreibung für den Einsatz von Stichproben, inklusive der Abgrenzung der Grundgesamtheit sowie insbesondere der Wahl von Auswahlverfahren und -Methoden für unterschiedliche Prüffelder. Darüber hinaus werden angrenzende Themenkomplexe, wie die Hinzuziehung von Experten und der Reviewprozess von Stichprobenprüfungen, dargestellt.[163]

Wie auch die IDW-Verlautbarungen erkennt der Audit Guide inhaltlich den hohen Stellenwert der Festlegung von den Stichprobenumfang beeinflussenden Parametern, namentlich Stichprobenrisiko, zu erreichende Prüfungssicherheit, Fehlererwartung sowie Fehlertoleranz, an. Neben der Beschreibung der Wirkungsrichtung der einzelnen Komponenten leitet der Audit Guide jedoch auch dazu an, wie derlei Parameter festgelegt werden, welche Bandbreiten für die Praxis typisch sind und wie Richtlinien zur Bemessung des Stichprobenumfangs entwickelt werden können.[164] Dies geschieht ähnlich wie in den F&A des IDW jeweils getrennt für Funktionsprüfungen des IKS sowie aussagebezogene Prüfungshandlungen. Die inhaltliche Tiefe geht dabei über die der F&A deutlich hinaus. Ergänzend wird eine umfangreiche Fallstudie für den gesamten Prozess der Auswahlprüfung bei Anwendung nichtstatistischer Stichprobenverfahren vorgestellt.[165]

Wie auch in den Veröffentlichungen von IDW und IFAC wird das Monetary Unit Sampling, nebst einer weiteren Fallstudie sowie quantitativer Überlegungen zu notwendigen Stichprobenumfängen und erreichbarer Prüfungssicherheit, gesondert betrachtet. Sowohl zur wertproportionalen Auswahl, als auch zur Durchführung von Attributstichproben und Sequentialtestverfahren für Funktionsprüfungen des IKS enthält der Anhang des Audit Guides Hilfstabellen, die eine heuristische Ermittlung notwendiger Stichprobenumfänge sowie die Bewertung der erreichten Prüfungssicherheit unter Berücksichtigung der entdeckten Fehler und des Prüfungsumfangs ermöglichen.[166]

Da die Anwendung von Strichprobenverfahren grundsätzlich die Verwendung unterstützender Software voraussetzt, enthält der Audit Guide zudem eine ergänzende Fallstudie zur Stichprobenerhebung bei Einsatz von IT-Software. Noch weiter gehen diesbezüglich

[163] Vgl. AICPA, a.a.O. 2014, Kapitel 3, 4.
[164] Vgl. AICPA, a.a.O. 2014, Rn. 37-65.
[165] Vgl. AICPA, a.a.O. 2014, Kap. 5.
[166] Vgl. AICPA, a.a.O. 2014, Anhang A bis D.

die ergänzenden „Technical Notes on the AICPA Audit Guide Audit Sampling“, welche u. a. entsprechende Formeln zur Programmierung von Makros in VBA[167] enthalten.[168]

Neben der Darstellung der genannten Verfahren beinhaltet der Audit Guide – anders als ISA 530 und IDW PS 310, jedoch ähnlich der ehemaligen HFA Stellungnahme 1/1988 – ein Kapitel zur klassischen Variablen-Stichprobe. Zu diesen Verfahren gehören neben der einfachen Mittelwertschätzung die gebundene Hochrechnung in Form der Verhältnis-, Differenzen- und Regressionsschätzung. Aufgrund der Tatsache, dass diese Verfahren nicht universell einsetzbar sind und ein effizienter Einsatz regelmäßig erst durch mehrfache Schichtung der Grundgesamtheit zum Tragen kommt, kann ihnen in praxi zunehmend jedoch lediglich geringe Relevanz beigemessen werden.

5.5 Technische Durchführbarkeit von Stichproben als Anwendungserschwernis

Es erscheint im Sinne der Eigenverantwortung des Abschlussprüfers begrüßenswert, die bewusste Auswahl als ein zur Stichprobe gleichwertiges Erhebungsverfahren zu klassifizieren, birgt eine ausschließliche Fokussierung auf repräsentative Verfahren neben den genannten Vorteilen und trotz der geschilderten unterstützenden Sekundärliteratur doch gleichermaßen veritable Risiken. Abseits aller normativen Betrachtungen besteht vor allem die Notwendigkeit der technischen Durchführbarkeit von Stichproben, d. h. entsprechend verwertbarer Buchhaltungsdaten.[169] Dabei können die Rahmenbedingungen mangelhaft sein, ohne dass dies dem Prüfer bewusst ist. Um Formen der digitalen Datenanalyse – deren Bestandteil auch Stichproben darstellen[170] – anzuwenden, benötigt der Prüfer stets und zwingend eine vollständige und auswertbare Datengrundlage. Als potenzielle Problemfelder, die in den Normen nur eingeschränkte Berücksichtigung finden, lassen sich exemplarisch folgende Sachverhalte anführen:

- Fehlerhafte Abgrenzung von Daten: die Grundgesamtheit enthält Elemente, die nicht der erwarteten Definition möglicher Auswahlelemente entsprechen,
- Unvollständige Datenbasis: es fehlen Datensätze aufgrund unbewusster (Bedienfehler) oder bewusster (dolose Handlungen) Nichterhebung,

[167] Visual Basic for Applications ist Bestandteil der Microsoft Office Anwendungen.
[168] Vgl. AICPA, a.a.O. 2014, Anhang 6; AICPA, Technical Notes on the AICPA Audit Guide, Audit Sampling, 2012.
[169] Vgl. Krahel/Titera, Accounting Horizons 2015, S. 409 ff.
[170] Vgl. z. B. Krehl, in: Deggendorfer Forum zur digitalen Datenanalyse e. V. (Hrsg.), Prozessoptimierung mit digitaler Datenanalyse, Bielefeld 2015, S. 113 ff.

- Redundanz: ein mehrfaches Vorkommen von Datensätzen verzerrt die Wahrscheinlichkeit der Ziehung einzelner Elemente,
- Mehrdeutigkeit: Datensätze enthalten mehr als ein potenzielles Auswahlelement (z. B. Rechnungen, die sich auf mehrere Lieferscheine beziehen oder Zahlungsavis für Sammelaufträge),
- Missing Values: die Datenbasis enthält zwar die vollständig abgegrenzte Menge potenzieller Auswahlelemente, es fehlen jedoch Werte in einzelnen relevanten Datenfeldern.

Der Prüfer muss sich dieser Einschränkungen stets bewusst sein, da ihr Vorliegen insbesondere die Effektivität repräsentativer Auswahlverfahren einschränken und ihn zu einem fehlerhaften Urteil führen kann. Aufgrund der Anforderungen durch die Finanzverwaltung sollte eine elektronisch auswertbare Datenbasis des gesamten Buchungsstoffes eines Geschäftsjahres grundsätzlich auch bei kleinen und mittelständischen Unternehmen ohne integrierte ERP-Systeme gegeben sein. Jedoch können auch GDPdU-Daten, die in der Regel ohne umfangreiche Aufbereitung durch Prüfungstools wie ACL[171] oder IDEA[172] bewältigt werden können, die genannten Mängel aufweisen.

5.6 Zusammenfassung

Die Übernahme von ISA 500 (rev.) und ISA 530 in nationale Prüfungsstandards erhöht den Stellenwert der statistischen Stichprobe, ohne jedoch im gleichen Zug die bewusste Auswahl zur vermeintlich schwächeren Alternative zu degradieren. Dies mag in praxi ggf. zwar leichte Veränderungen im Rahmen der Durchführung der Abschlussprüfung mit sich bringen, eine Einschränkung der Handlungsfreiheit des Prüfers durch IDW PS 310 kann indes nicht konstatiert werden. So hat der Abschlussprüfer auch weiterhin die Möglichkeit, bei Vorliegen entsprechender Voraussetzungen primär auf sein prüferisches Ermessen zu vertrauen und auf eine selektive Auswahl der zu prüfenden Geschäftsvorfälle zurückzugreifen (vgl. Tabelle 5.2).

[171] Audit Command Language, Prüfsoftware der Fa. ACL Services Ltd.
[172] Interactive Data Extraction and Analysis, Prüfsoftware der Fa. Audicon GmbH.

Tabelle 5.2: Empfohlene Auswahlverfahren in Abhängigkeit des Prüffelds

	Empfohlene Auswahlmethode/Statistisches Verfahren				
Prüffeld	Ausschließlich Vollerhebung	Bewusste Auswahl	Zufallsimitierende Auswahl	Echte/unechte Zufallsauswahl	Wertproportionale Auswahl (MUS)
Internes Kontrollsystem: Funktionsprüfungen	X	O	--	X	
Anlagevermögen (Zugänge)		X	--	X	X
Anlagevermögen (Abgänge)		X	--	X	
Vorratsvermögen (Inventur, Mengengerüst)		O	X	X	X
Vorratsvermögen (Bewertung, Preistest)		O	--	X	X
Forderungen (Saldenbestätigungen)		X	--	X	X
Liquide Mittel (Bankbestätigungen)	X				
Personalrückstellungen		O	--	X	
Verbindlichkeiten (Saldenbestätigungen)		O	--	X	
Umsatzerlöse (Realisierung)		O	--	X	X
Umsatzerlöse (Periodenabgrenzung)		O	--	X	X
Aufwand (Periodenabgrenzung)		X	--	X	

X = uneingeschränkt anwendbar
O = Anwendung in Kombination mit Zufallskomponente empfohlen
-- = Anwendung nicht empfohlen

Für den Fall, dass er eine repräsentative Auswahl als Basis zur Einholung von Prüfungsnachweisen vorzieht, erhält er mit IDW PS 310 jedoch nun einen im Vergleich zur HFA Stellungnahme 1/1988 einfacher verständlichen und weithin „schlanken“ Standard. Aufgrund der inhaltlichen Deckungsgleichheit zu ISA 530, welcher in weiten Teilen durch die AU Section 530 des AICPA motiviert ist, kann der Prüfer mit einer Kombination aus dem AICPA Audit Guide „Audit Sampling“ und den F&A des IDW zugleich auf umfangreiche Hilfestellungen zurückgreifen. Dabei bietet es sich geradezu an, insbesondere die angeführten, im Audit Guide enthaltene Dokumentation in Form von Hilfstabellen zur Ermittlung der generierten Prüfungssicherheit sowie der Bestimmung eines angemessenen Stichprobenumfangs als Ergänzung zum weniger statistisch ausgelegten IDW PS 310 mitsamt der F&A zu Rate zu ziehen. So enthalten die Veröffentlichungen des AICPA zahlreiche „Decision Aids“, welche die praktische Anwendung von Stichproben gegebenenfalls deutlich erleichtern können. In Kombination erhält der geneigte Anwender damit ein praktikables Handwerkszeug, welches zur Wahl angemessener Erhebungs- und Auswertungsverfahren sowie deren Durchführung in vielerlei Prüfungssituationen ausreichen wird. Ungeachtet jener Hilfestellungen gilt es sich dabei jedoch stets der Begrenzungen jedweder Anwendung von Stichproben in Form der aufgezählten Problemfelder zu vergegenwärtigen. Denn allein das fundierte Wissen um das Für und Wider der einzelnen

Auswahlmethoden sowie deren einhergehenden Anwendungsgrenzen und -konsequenzen versetzt den Abschlussprüfer letztlich in die Lage, das Instrumentarium stichprobengestützter Auswahlmechanismen auch in praxi valide und reliabel einzusetzen.

6 Does (Sample) Size Matter? The Sensitivity of Auditors to a Revision of Non-Statistical Audit Sampling Standards

6.1 Publication Details

Abstract: Non-statistical sampling is a common technique employed by financial statement auditors to determine sample sizes. Auditors' judgments in this context are important, because improper sample size appraisals can lead to increased sampling risk, and potentially the issuing of an inappropriate audit opinion. This paper examines the effect of a semi-structured aid for auditors in sample size decision making in non-statistical sampling. The aid we examine, which draws on ISA 530 guidance, encourages a structured approach by combining the individual effects of various sampling plan properties on sample size. Drawing on the anchoring and adjustment heuristic, we hypothesize that the decision aid improves auditors' sample size decisions by reducing their cognitive strains. We test our hypothesis with an experiment among 179 German auditors. We find that auditors exposed to our decision aid recommend larger samples. Our results also indicate that auditors are generally aware of the factors that impact sample size, but sometimes make insufficient adjustments in high audit risk settings.

Co-Authors: Prof. Dr. Anna Gold, Prof. Dr. Christiane Pott.

Keywords: audit sampling, sample size, auditor judgment, anchoring and adjustment, decision aids.

JEL-Code: C91, M42, D91.

Publication Status: Working paper. This paper was presented at the 24th Annual International Symposium on Audit Research (ISAR 2018) in Maastricht (The Netherlands), and at the 2018 Annual Meeting of the American Accounting Association in Washington D.C. (USA).

6.2 Introduction

In this study, we examine the influence of ISA 530 guidance on auditors' non-statistical sample size judgments. While statistical sampling dominated audit practice in recent decades (Elder et al. 2013), today non-statistical sampling is common in the field (Christensen et al. 2015). In non-statistical sampling, sample size judgments are much more prone to decision biases owing to the subjective assessment of sample size input parameters (Christensen et al. 2015; Elder et al. 2013). ISA 530 offers guidance in this context as a semi-structured decision aid which accentuates the combination of individually important sample size input factors in the application of non-statistical sampling plans. The standard will become mandatory in German audits from 2018. These developments raise the question of whether more guidance leads to changes in auditors' sample size decisions and consistency of judgment. Examining the effects of decision aids on auditor behavior is important, as previous research has shown that such guidance can have positive implications for audit practice (Messier et al. 2001) but might also yield unintended outcomes (Fay et al. 2015). Insights into the current use of audit sampling and the sensitivity to changes in audit standards are limited (Elder et al. 2013). We are not aware of any studies to date which empirically investigate the effect of semi-structured guidance on auditors' non-statistical sample size decisions.

Some researchers claim that audit standards have not kept pace with the development of big data environments (Krahel and Titera 2015). However, almost all audits still involve sampling to gather substantive audit evidence (Sibelman 2014). Current auditing standards require the auditor to gather sufficient and appropriate audit evidence (AICPA 2011; IAASB 2009; PCAOB 2014). While appropriateness addresses relevance and reliability, sufficiency is determined by the extent of testing (ISA 530.7, A10), commonly defined by the chosen sample size. Auditing firms allow the use of various sampling techniques and procedures, including statistical as well as non-statistical sampling (Christensen et al. 2015). Professional acceptance and cost efficiency measures have led less restrictive non-statistical sampling plans to be favored by auditors in the field (Christensen et al. 2015; Maingot and Quon 2009; Schwartz 1997). Unfortunately, several elements of non-statistical sampling plans are open to subjective assessment and provide more scope for the violation of statistical principles, including a potentially insufficient amount of testing. The auditor faces two conflicting objectives: (1) maximizing profits and (2) minimizing audit risk (Swearingen and Hansen 1990). Biases due to decisions within the optimization

of an auditors' utility function include the avoidance of extensive testing. However, these biases can lead to potentially compromised auditing effectiveness owing to underestimated risk and a lack of required assurance (Elder and Allen 2003).

Setting sample size based on one's judgment is a complex task, as multiple and sometimes conflicting parameters need to be considered. Early research in auditing (Kachelmeier and Messier 1990; Messier et al. 2001; Elder and Allen 2003) and psychology (Tversky and Kahneman 1971, 1974; Kahneman and Tversky 1972) shows that auditors reduce this complexity by relying on a limited number of heuristics that simplify certain judgmental operations but increase the risk of severe and systematic errors (Tversky and Kahneman 1974). As opposed to the former normative status in Germany, the revised standard now includes qualitative ISA 530-based guidance for the determination of judgmental sample sizes when using non-statistical sampling (ISA 530 Appendix 3). Pertinent to the current study, this guidance constitutes a semi-structured decision aid which is supposed to help auditors evaluate appropriate and consistent sample sizes, suited to individual auditing situations.[173]

We examine the impact of two key factors that are explicitly considered in the wording of the ISA 530 guidance – the risk of material misstatement, and audit assurance by additional testing procedures – as these factors are often less resolutely acknowledged by auditors than they should be (Mauldin and Wolfe 2014; Durney et al. 2014; Sibelman 2014). We test our research question using a 2 x 2 x 2 between-subjects experiment, in which we manipulate the extent of ISA 530-guidance (absence vs. presence) as well as different levels of inherent and control risk (high vs. low) and the number of additional audit procedures performed (absence vs. presence). A total of 179 practicing German auditors participated in an experiment in which they were asked to determine sample sizes for a given case using common non-statistical sampling. ISA 530 will become mandatory in German audits from 2018, raising the question of whether more guidance leads to changes in auditors' sample size decision making in an environment where auditors are not yet affected by such guidance. We predict, and find, that the presence of the decision aid significantly increases participants' sample size determination. However, auditors in any guidance condition tend to underestimate how much substantive testing is needed in

[173] As categorized by Messier (1995), ISA 530 provides a simple but universally usable decision aid, rather than a complex expert system. Simple decision aids help the decision maker primarily in considering and combining relevant information while lowering his or her cognitive effort.

high-risk environments with no further testing procedures performed, failing to choose satisfactory sample sizes as compared to statistically derived sample sizes. In line with previous research on decision aids, we provide evidence that ISA 530 counteracts these flaws by refocusing auditors' attention onto the most important factors impacting sample size. Interestingly, additional results indicate that participants frequently underestimated the necessity of large samples, even in the presence of the ISA 530 guidance.

Our study contributes to auditing theory and practice in three key ways. First, we examine whether additional guidance on judgmental sample size determinations changes auditors' behavior towards more consistent and risk-adjusted decisions, potentially achieving higher auditing quality. Second, of importance to practice and standard-setters, we investigate the effect of a recently changed auditing standard. We test our hypotheses for various types of auditing situations, including different levels of inherent and control risk as well as the number of further substantive testing procedures performed, to estimate the efficacy of the change in standards. To the best of our knowledge, no prior study has tested the implication of ISA 530 guidance on auditor behavior in assessing non-statistical sample sizes.

Third, we offer important practical suggestions for researchers, educators, standard-setters, regulators and practitioners, as we provide a better understanding of how auditors apply their sampling plans, and an assessment of auditors' relative performance in various auditing settings. As auditing efficacy and the proper documentation of judgmental sampling decisions are consistently reviewed by inspections and internal quality reviews, the examination of sample size decisions is of huge importance. Findings may help advise standard-setters of necessary improvements to current auditing standards. The results may also be useful in auditing practice and help find starting points for the establishment of improved decision aids. Furthermore, our study reveals areas of potential future research and points out inevitable areas for improvement in auditing education, which may help prevent the continuation of disclosed concerns and better understand the application of audit sampling theory in current auditing environments.

The remainder of this paper is organized as follows. The second section discusses sampling and sample size decisions in auditing contexts, reviews prior literature on this topic, and explains the development of our hypotheses. The third section describes our research design and data collection. The results and implications are discussed in the fourth section. The final section summarizes the results and discusses implications for study.

6.3 Background and Hypothesis Development

6.3.1 Sample Size and Critical Determinants in Non-Statistical Audit Sampling

Assessing probabilities and predicting values are complex tasks. To reduce this complexity to simpler judgmental operations, individuals often rely on heuristics. In this approach humans are prone to decision biases, obtaining different outcomes than statistical foundation might yield. A consistent finding in research into human information processing is that decision makers often reach judgments that deviate from values expected from normative models. A common heuristic, which explains inaccurate sample size judgment, is that of anchoring and adjustment (Fay et al. 2015; Kowalczyk and Wolfe 1998; Libby 1985; Tversky and Kahneman 1974).[174] When anchoring is used, an initial estimate is followed by an insufficient adjustment as more information is received and considered. In the audit sampling context, this refers to common situations where various information accumulates, e.g. prior years' experience, allocated inherent and control risks, the extent of additional testing procedures towards the same auditing objective, and many more. Insufficient adjustments or a poor initial estimate can be the result of inadequate processing of this information (Butler 1986; Joyce and Biddle 1981). Fay et al. (2015) find evidence that auditors focus on generally known firm-standard procedures when planning and performing audit procedures, which leads to costly over-auditing or to under-auditing that significantly increases sampling risk. Joyce and Biddle (1981) find that the establishment of initial values is influenced primarily by the formulation of an audit problem and the individual's experience. In line with others, Ashton (1990), Kleinmuntz (1990), and Boatsman et al. (1997) find that auditors are regularly distrustful of decision aids and overly optimistic regarding intuitive decisions.

Auditing standards recommend factors to be considered in the general application and arrangement of sampling plans as well as in determining sample size (Christensen et al. 2015; Houghton and Fogarty 1991). While ISA 530 does not provide quantitative measures, it does enumerate relevant characteristics regarding the client and the auditing objective, accompanied by the effective direction of these characteristics on sample

[174] We focus on the anchoring heuristic, as auditors rely heavily on experience gathered in other audits and make inappropriate adjustments to client-specific properties. This behavior is induced particularly by force of habit, resulting in a 'same-as-last-year' approach.

size.[175] In this study, we view the ISA 530 guidance as a decision aid with the potential of improving audit judgment quality (see Table 6.1 for a summary of the guidance).[176]

Table 6.1: Examples of Factors Influencing Sample Size for Tests of Details

Factor	Effect on Sample Size
1. An increase in the auditor's assessment of the risk of material misstatement	Increase
2. An increase in the use of other substantive procedures directed at the same assertion	Decrease
3. An increase in the auditor's desired level of assurance that tolerable misstatement is not exceeded by actual misstatement in the population	Increase
4. An increase in tolerable misstatement	Increase
5. An increase in the amount of misstatement the auditor expects to find in the population	Increase
6. Stratification of the population when appropriate	Decrease
7. The number of sampling units in the population	Negligible

To arrive at valid audit assertions, it is crucial that one maintain consistent sample size decisions under similar conditions. Ideally, non-statistical audit sampling should approximate statistical approaches (Hall et al. 2012, Maingot and Quon 2009; Houghton and Fogarty 1991). By contrast, Christensen et al. (2015), as well as Swearingen and Hansen (1990), found non-statistical approaches to dominate statistical approaches owing to the substantially smaller sample sizes used. They attribute this finding to the fact that the components to be considered in setting sample sizes judgmentally are comprehensive and help auditors circumvent normative requirements. Theory suggests that even if the likelihood of misjudgment in each component is slight, there is a high probability of a flawed aggregate decision (Tversky and Kahneman 1974). Thus, debiasing techniques such as the ISA 530 decision aid could be useful in mitigating such adverse consequences.

6.3.2 Literature Review

Mock and Turner (1981) suggest that changes in internal control and normative guidance have a real effect on auditors' sample size decision making. While their assumptions were made by the construction of a prototypical simulation, Kachelmeier and Messier (1990) were the first to experimentally assess whether auditors determine sample sizes in accordance with auditing standards and statistical theory. Results not only indicated a tendency of auditors to choose sample sizes smaller than statistical theory would advise but they

[175] Similar advice can also be found in the AICPA auditing guide "Audit Sampling" (AICPA 2017).
[176] The complete aid is included in the final instrument of our experiment, in the Appendix.

also found that auditors "work backward", assessing risk levels ex post to achieve a desired sample size.

In a follow-up experiment with 149 experienced auditors, Messier et al. (2001) examined whether sample sizes were consistent with the then-revised AICPA audit sampling guide. Alarmingly, average sample sizes diminished significantly as compared to the previous study, leading to rising concern about auditors' judgments and exposed audit risk. Similar results were achieved in an archival study by Elder and Allen (2003). Much later, Trompeter and Wright (2010) interviewed 36 auditors and found that even when technological advancement was prevalent, auditors still favored tests of details over additional audit procedures. However, participants seemed to overestimate the marginal gains of substantive testing procedures. A PCAOB inspection of 1,662 audits for the years 2004 to 2007 included general concerns on sample selection, as well as sample sizes being too small (PCAOB 2008, 2014).

Durney et al. (2014) found larger average sample sizes than Elder and Allen (2003) but did not report the sensitivity of sample sizes to clients' risk. Addressing this, a case-based experiment by Mauldin and Wolfe (2014) required 81 Big 4 seniors to map control deficiencies into modifications of substantive tests of details. Results indicated that a straightforward modification of the extent of substantive audit evidence was not well established, as a significant number of auditors failed to adequately adjust the extent of substantive tests of details when control deficiencies were present.

Most recently, in a survey of audit firms' policy departments, Christensen et al. (2015) examined the sampling policies of audit firms. Even though all the firms surveyed basically followed the same auditing standards, sampling policies differed greatly, which resulted in a noticeable variance in the nature and extent of audit sampling in the field.

To summarize, prior research has indicated a strong tendency towards insufficient sample size decisions by auditors. We therefore aim to examine the actual effect of a semi-structured decision aid provided by ISA 530 on auditors' sample size decisions.

6.3.3 Hypothesis Development

Both archival and experimental research provide evidence that auditors' sample size judgments are not flawless. Questions arise as to if and how certain individual characteristics of an audit objective, such as client-specific risk and the overall audit strategy, correlate with those judgments, as the observed concerns are in direct conflict with normative

requirements. Theory suggests that misconduct can be partially attributed to the high cognitive complexity due to the multifactorial alignment of sample size inputs (Butler 1986; Joyce and Biddle 1981). According to Fischoff (1982), biased judgments can occur through (1) faulty tasks, (2) faulty judges or (3) a judge-task mismatch. Addressing (2) only, we constructed a job-related task focusing on a rather common audit procedure. Restructuring the underlying decision-making process provides an opportunity to mitigate the adverse consequences of auditors' insensitivity to certain features of the task. With respect to theory and prior research, the following hypotheses were derived.

Results from Tversky and Kahneman (1971, 1974) and Kahneman and Tversky (1972) indicate that individuals expect small samples to provide considerably more information than statistical theory justifies. This "belief in the law of small numbers" occurs when sample sizes are small or planned (Tversky and Kahneman 1971), yielding a cognitive bias in which individuals are not aware of the extent to which statistical power will diminish as sample sizes fall. Cost and workload considerations of audit applications might be of concern, not providing the desired assurance for one to accept an accounting population. The decision aid renews the focus of auditors on every single parameter. In line with Rohrmann (1986), we expect that the aid might (1) increase the perceived importance of judgments, (2) improve the structure of the input data, (3) lead to a justification of decision aid outputs, (4) increase or reduce judgment consistency, (5) allow a user to circumvent its intended use. While Kachelmeier and Messier (1990) addressed (1), (2), and (5), we refer to (1), (2), and (4). We expect that individuals without the decision aid will intuitively recommend smaller sample sizes, as bypassing factors suggesting larger samples is more likely when no justification for the bypass strategy is required. On the other hand, ISA 530 breaks down the process and makes auditors aware of the initial input parameters for a qualitative sample size decision. As ISA 530 does not provide quantitative sample size recommendations, (3) and (5) cannot be addressed directly. As a result, we formulate our first hypothesis:

H1a: Participants provided with ISA 530 Appendix 3 will recommend larger sample sizes than participants making intuitive sample size decisions.

Our second hypothesis addresses the consistency of auditors' sample size judgments. Previous research in auditing has found contrary indications towards this consideration. Ashton and Willingham (1988), as well as Ashton (1983) and Elliott (1983), propose judgment dispersion to be smaller when decision aids are used. However, Jiambalvo and

Waller (1984) argue that the splitting up of a complex problem into less complex individual decisions does not necessarily lead to a higher degree of accuracy that will increase dispersion. As a matter of routine auditors expect a certain outcome from diagnostic judgments to confirm their hypotheses (Smith and Kida 1991). In keeping with anchoring theory, Butler (1986) and Fay et al. (2015) find that auditors judgmentally evaluate samples based on expectations resulting from personal experience. ISA 530 does not lead to quantitative sample size measurements, but rather demands a qualitative assessment of sampling risk when non-statistical procedures are used. We expect auditors' sample size decisions to be exposed to the anchoring effect, with the decision aid mitigating this effect as awareness of the individual factors influencing sample size drives auditors to deviate from intuitive judgments:

H1b: Recommended sample sizes in the ISA 530-guided group will have a higher dispersion than those in the intuitive group.

Bonner et al. (1996) find that "decomposition-and-mechanical-aggregation" aids can have a positive influence on auditor judgments if they successfully counteract mismatches between the organization of an audit task and that of the auditors' knowledge. With reference to the risk-adjusted audit approach, Christensen et al. (2015) indicate that an auditor's sample size judgment should be heavily focused on (1) the level of inherent and control risk and (2) the extent of additional testing procedures directed toward the same audit objective. In the field, sample sizes are sometimes only weakly associated with risk assessments (Mauldin and Wolfe 2014; Elder and Allen 2003). Furthermore, audit inspection reports indicate that auditors often fail to implement appropriate modifications to substantive tests of details when ineffective controls are discovered (PCAOB 2008). As ISA 530 directly addresses sample size sensitivity to risk assessments and the extent of additional substantive procedures that are performed, we hypothesize the following:

H2: Recommended sample sizes will be larger in high-risk environments as compared to low-risk environments, with the effect being strengthened when ISA 530 guidance is present.

H3: Recommended sample sizes will be lower when additional substantive testing procedures toward the same audit objective are performed, with the effect being strengthened when ISA 530 guidance is present.

6.4 Research Method

6.4.1 Participants

A total of 179 practicing auditors from nine audit firms in Germany participated in our experiment. Incentives were set by indicating that each attendance led to a donation of five euros to one of five charity organizations to be chosen from. As interviews during pretests indicated that auditors on any hierarchical level can generally be responsible for sample size judgments, we argue that these individuals were suitable for the task. All participants indicated at least a working knowledge of audit sampling and the routine use of non-statistical sampling in the field. However, the distribution of audit experience was heavily skewed. The participants had on average 4.9 years of audit experience (standard deviation 4.47), with 57 percent having less than 4 years' experience. Of all the participants, 23 percent had already prepared for the German Public Accountant's exam.

6.4.2 Research Design and Experimental Procedure

We asked partners of the firms to distribute the instrument to audit professionals with at least one year's experience. Random assignment of the instrument resulted in a balanced distribution of observations across experimental conditions. We conducted the experiment in mid-2017, when German auditors were probably not aware of the normative changes being examined, as these will come into effect in the audit season of 2018.
We assume ISA 530 Appendix 3 to be a semi-structured decision aid which decomposes a judgmental determination of sample sizes. The adoption of ISA 530 in Germany became effective for audits of financial statements for periods beginning on or after December 15th, 2016. We focus on two common factors that are crucial to any sample size assessment and that are explicitly considered in the wording of the ISA 530 guidance: the level of inherent and control risk, and the extent of additional substantive testing procedures performed.

The experimental task was a modified version of an instrument developed by Kachelmeier and Messier (1990). We assume that nowadays accounts receivable is the most common audit field to provide a basis for our experiment, because of its reasonable homogeneity and the uniform audit procedures performed in the field (Elder and Allen 1998). We defined the population of interest to be an accounts receivable balance represented by 978 individual debt accounts, with single debt accounts ranging between €1,000

and € 749,000. Participants were provided with general information on the case. This included a narrative about the company, unaudited balance sheets and income statements, information on the efficacy of internal controls, materiality considerations, and the extent of additional substantive testing procedures that were performed. The relevant information was complemented by various distracter cues. It was then assumed that the manager in charge was determined to draw a non-statistical random sample from the population. As sample size judgments vary largely when the underlying population is highly skewed, we predefined that a stratification of the population leads to the selection of key items comprising the six largest debt accounts (complete-count stratum). The participants' task was to draw an additional sample of *n* items from the remainder of the population, represented by 972 debt balances (selective stratum). Participants were told that the sample size in the selective stratum should preferably be kept small because of the expected quantity of alternative audit procedures. In line with findings from Christensen et al. (2015), who found that usually not more than 100 sample items are drawn in the field, participants were told not to request more than a maximum of an additional 100 confirmations.

A 2 x 2 x 2 between-subjects factorial design with fixed factors was used to test our hypotheses. We manipulated the variable of interest – presence of *GUIDANCE* – between subjects. Participants were randomly assigned to the two *RISK* conditions with either reliable or unreliable internal controls.[177] Similarly, *PROCEDURES* was manipulated at two levels, either with or without further substantive tests being performed. All experimental factors were fully crossed, resulting in eight experimental cells. The instrument was pre-tested in a pilot study. Five academics and six auditors (three seniors, two managers, one partner) completed the study in an average of 46 minutes. A debriefing of the pilot participants revealed that the task and manipulations were realistic and understandable. Several suggestions were made to further improve the realism of the task. All suggestions were incorporated into the final instrument. The average time taken by the participants to complete the final instrument was 31 minutes.

6.4.3 Model

The following linear regression model was used to test the hypothesis:

[177] The stronger condition as compared to the weaker condition was determined by a larger accounting department, an adequate substitution regulation, no revealed errors in tests of internal control and well implemented new controls as a reaction towards prior year findings.

SAMPLE SIZE = f(*GUIDANCE, RISK, PROCEDURES, AUDIT EXPERIENCE, TASK-RELATED EXPERIENCE, KNOW_STAT, BIG 4, EXAM_PREP*)

Table 6.2: Model Predictions and Description of Variable Coding

Variable	Predicted Sign	Coding description
Dependent Variable:		
SAMPLE SIZE		= 0 to 100 rating in steps of 10 [0; 1 to 10; 11 to 20; … ; 91 to 100]
Independent Variables:		
GUIDANCE	+	= 1 if ISA 530 provided = 0 otherwise
RISK	+	= 1 if reliable = 0 otherwise
PROCEDURES	-	= 1 if additional tests performed = 0 otherwise
Control Variables:		
AUDIT EXPERIENCE	+	= Years of Experience in Auditing
TASK RELATED EXPERIENCE	+	= 1 to 7 rating (where 1 = none and 7 = extensive)
KNOW_STAT	+	= 1 to 3 rating (where 1 = none and 3 = extensive)
BIG 4	+/-	= 1 if Big 4 member = 0 otherwise
EXAM_PREP	+	= 1 if Auditor prepared for Exam = 0 otherwise

6.4.4 Dependent Variable

Our aim was to examine the effect of the presence of decision aids on the extent of substantive audit evidence gathered by auditors. Our dependent variable was therefore the participants' assessment of the *SAMPLE SIZE*. We assume that participants' decisions about sample size are made using estimates of likelihood that reflect any information a participant considers relevant. As Kachelmeier and Messier (1990) found that participants generate implausible outliers, *SAMPLE SIZE* in our model was coded from 0 to 100, in intervals of 10 (0 to 10; 11 to 20;…; 91 to 100).

6.4.5 Independent Variables

GUIDANCE. Elder et al. (2013) state that auditing standards play a central role in auditors' decisions, but the presence of decision aids has been found to have contrary effects (Ashton 1983; Elliott 1983; Mock and Turner 1981; Jiambalvo and Waller 1984). As ISA 530 formally acknowledges non-statistical sample sizes, we expect the availability of this guidance to have a significant effect on auditors' judgments. Participants in the guidance condition received the full wording of the corresponding ISA 530 passages, while the rest of the narrative was held constant. With respect to the anchoring heuristic, we expect the presence of ISA 530 to increase the magnitude of sample sizes as well as their dispersion.

RISK. Previous research has indicated that auditors are more likely to modify their testing approaches as the engagement risk increases (Fay et al. 2015; Cohen and Kida 1989), while audit programs generally do not have the necessary adjustment sensitivity for inherent and control risk (Mock and Wright 1999; Walo 1995). To test for this effect, *RISK* was manipulated at two levels by altering the case narrative, comprising a "weak to fair" and a "fair to strong" condition (Mock and Turner 1981). This procedure was also employed by Kachelmeier and Messier (1990). However, the wording in our case was altered toward a slightly stronger distinction between the conditions.

PROCEDURES. In their case, Kachelmeier and Messier (1990) included, but did not manipulate, a certain level of additional substantive testing procedures. The ISA 530 decision aid directly addresses the impact of additional substantive tests on sample size. As theory suggests that auditors should be intuitively aware of the effect, we incorporated this factor into our design. The variable was manipulated on two levels. One condition implied that no further substantive tests were executed. In the other condition, it was asserted that complementary audit work had been scheduled, including analytical audit procedures as well as revenue recognition and year-end cut-off procedures.

6.4.6 Control Variables

AUDIT EXPERIENCE. According to Taylor and Dunnette (1974), as well as Libby (1985), experience is important in a variety of decision contexts. Experienced auditors may have seen a lot of audit objectives, such as in our case. However, contrary to this expectation, previous research has found that experienced auditors tend to use internal anchors, even when auditing new clients (Fay et al. 2015; Bedard and Johnstone 2010;

Mock and Wright 1999). Auditing experience is measured by the participants' declaration of their experience in years.

TASK-RELATED EXPERIENCE. In addition to general auditing experience, we control for experience with non-statistical sampling, as auditors with several years' experience do not necessarily apply these procedures regularly. Hence, participants made an indication of their frequency of dealing with similar audit tasks in the field, measured on a seven-point Likert scale, ranging from zero (never) to seven (often).

KNOW_STAT. Grounded knowledge of statistical coherences favorably influences subjective probability assessment (Libby 1985; Dubé-Rioux and Russo 1988). In line with Winkler (1967a) and Stael von Holstein (1972), we expect that greater statistical understanding will favor a consistency of response with statistical principles, especially when no additional guidance is provided. With wording slightly altered to accord with Blocher and Bylinski (1985), we measure statistical knowledge by participants' self-assessment on three levels: no skills, moderate skills, and extensive skills.

BIG 4. Individual differences between the Big 4 aside (Christensen et al. 2015), global sampling policies of these firms may already be built upon IAASB announcements, leading to corresponding participants being implicitly aware of the tested decision aid. Because of anonymity constraints, we distinguished only between Big 4 and non-Big 4 firms by way of participants' indications in the additional questionnaire.

EXAM_PREP. Throughout the examination courses for the auditor exam, normative prescription and statistical theory are discussed in great depth. Although previous studies examining auditors' sampling behavior have controlled for experience, we argue that because of the complexity of the underlying statistical concepts, only participants who have already prepared for the auditing exam are fully aware of the consequences of each factor to be considered. Hence, we expect participants with this background to make sample size judgments that are more accurate and comparable to statistical theory.

6.5 Results

6.5.1 Manipulation Checks

Prior to data analysis, we ensured that participants appreciated the manipulations by having them complete manipulation checks within the debriefing questionnaire. Of all the participants, 98 percent correctly indicated the presence of additional sample size

guidance in the guidance-present condition, indicating a successful manipulation of *GUIDANCE*. Perceived strength of internal control was rated on a 7-point scale with 1 denoting not reliable at all and 7 denoting very reliable. The mean (standard deviation) rating for *RISK* in the low-risk environment was 2.72 (1.39), while in the high-risk environment it was 5.84 (1.20). The difference is statistically significant ($t = 16.10$; $p < 0.000$), and indicates a successful manipulation of *RISK*. Similarly, the participants were asked to assess the perceived extent of additional substantive testing procedures on a 7-point scale ranging from 1 (none) to 7 (many). The mean (standard deviation) rating for *PROCEDURES* was 2.33 (1.69) in the no-tests condition, and 5.42 (1.45) in the further-tests condition. The difference is statistically significant at ($t = 13.14$; $p < 0.000$), indicating a successful manipulation of *PROCEDURES*.

6.5.2 Descriptive Statistics

Descriptive statistics for the eight treatment conditions are shown in Table 6.3. Panel A shows cell means and standard deviations for the dependent variable *SAMPLE SIZE*. Panel B of Table 6.3 shows means and standard deviations by experimental condition. When *RISK* is high, *SAMPLE SIZE* means are considerably larger than in low-risk conditions. *SAMPLE SIZE* means are also larger when *PROCEDURES* are absent, as compared to conditions with further tests being performed. Moreover, participants' *SAMPLE SIZE* means are larger for any combination of *RISK* and *PROCEDURES* when *GUIDANCE* is present, except when *RISK* is high, and *PROCEDURES* are absent. Standard deviations are comparable in all conditions, especially for the manipulation of *GUIDANCE*, as shown in Panel B of Table 6.3.

6.5.3 Main Results

The Kolmogorov-Smirnov-Test demonstrates that a normal distribution of our data can be assumed ($p = .088$), with *SAMPLE SIZE* being the dependent variable and *GUIDANCE* the independent variable. The homogeneity of variances was asserted using Levenes' Test ($p = .9544$). We test our hypothesis using a 2 x 2 x 2 ANCOVA design with three between-subjects factors: *GUIDANCE*, *RISK*, and *PROCEDURES*. Panel A of Table 6.4 gives the full model results, including our control variables: *EXPERIENCE*, *TASK-RELATED EXPERIENCE*, *KNOW_STAT*, *BIG 4*, and *EXAM_PREP*.

Hypothesis H1a states that participants provided with the decision aid on non-statistical sample sizes, included in ISA 530, will recommend larger samples than participants not provided with this guidance. The main effect of *GUIDANCE* on *SAMPLE SIZE* is found to be significant, with a *p*-value of .0881. The main result we observe here is that an accurate judgment on the extent of testing is highly dependent on auditors' perceptions of the circumstances of the audit objective. When no normative guidance on sample size was given, participants recommended smaller samples, leaving more room for sampling risk. Thus, we find strong support for our H1a.

Hypothesis H1b was directed toward the consensus of participants' sample size recommendations. Panel A of Table 6.5 gives the results of an ANCOVA, with *SAMPLE* SIZE being the dependent variable and *GUIDANCE* the sole independent variable. We observe that variances in the ISA 530-guided condition and the intuitive condition are almost alike, hence the dispersion of *SAMPLE SIZE* in both conditions of *GUIDANCE* is similar. In agreement with H1a, the effect of *GUIDANCE* is significant with a p-value of 0.0820, revealing an effect of the decision aid on the extent of testing through larger samples. However, presence of the decision aid does not lower differences in opinions toward the correct sample size.

Table 6.3: Descriptive Statistics for *SAMPLE SIZE*[a]

Panel A: Descriptive Statistics by Cell (n = 179)

		GUIDANCE[d]	
PROCEDURES[b]	***RISK***[c]	**Present Mean (Std. Dev.)**	**Absent Mean (Std. Dev.)**
Present	High	41.86 (23.00)	29.50 (16.35)
	Low	22.31 (20.16)	18.15 (16.22)
Absent	High	50.70 (27.84)	52.22 (29.41)
	Low	33.68 (19.91)	25.50 (17.99)

Panel B: Descriptive Statistics by Condition

GUIDANCE		*RISK*		*PROCEDURES*	
Present Mean (Std. Dev.)	**Absent Mean (Std. Dev.)**	**High Mean (Std. Dev.)**	**Low Mean (Std. Dev.)**	**Present Mean (Std. Dev.)**	**Absent Mean (Std. Dev.)**
36.81 (24.88)	31.32 (24.31)	43.99 (26.05)	24.71 (19.17)	27.58 (20.91)	40.66 (26.51)
n = 90	n = 89	n = 87	n = 92	n = 90	n = 89

[a] *SAMPLE SIZE* is measured by the indicated number of sampling items.
[b] *PROCEDURES* is manipulated between subjects at two levels: Present and absent.
[c] *RISK* is manipulated between subjects at two levels: high-risk and low-risk.
[d] *GUIDANCE* is manipulated between subjects at two levels: Present and absent.

Table 6.4: Test of H1a, H2, H3

Panel A: Full Model[a] ANCOVA (*SAMPLE_SIZE*)

Source	SS	df	F	p-value
GUIDANCE	1,401.85	1	2.94	0.0881
RISK	13,254.25	1	27.83	0.0000
GUIDANCE x RISK	2.41	1	0.01	0.9434
PROCEDURES	6,645.11	1	13.95	0.0003
GUIDANCE x PROCEDURES	863.81	1	1.81	0.1799
RISK x PROCEDURES	556.56	1	1.17	0.2813
GUIDANCE x RISK x PROCEDURES	1,433.81	1	3.01	0.0846
TASK RELATED EXPERIENCE	632.13	1	1.33	0.2509
AUDIT EXPERIENCE	711.89	1	1.49	0.2232
BIG 4	669.06	1	1.40	0.2376
EXAM_PREP	1,326.21	1	2.78	0.0971
KNOW_STAT	521.57	1	1.10	0.2969

Panel B: Simple Effects of *GUIDANCE* at Different Levels of *RISK* and *PROCEDURES*

		Effect of *GUIDANCE*[b]	
PROCEDURES	***RISK***	**F**	**p-value**
Absent	High	5.43	0.0210
	Low	0.46	0.4993
Present	High	0.40	0.5260
	Low	1.04	0.3090

[a] Between-subjects tests are performed for *GUIDANCE*, *PROCEDURES and RISK*.
[b] Between-subjects tests are performed for *GUIDANCE*.
SAMPLE SIZE is measured by the indicated number of sampling items.
PROCEDURES is manipulated between subjects at two levels: Present and absent.
RISK is manipulated between subjects at two levels: high-risk and low-risk.
GUIDANCE is manipulated between subjects at two levels: Present and absent.
AUDIT EXPERIENCE is measured by the indicated years of experience in audit.
TASK RELATED EXPERIENCE is measured by the indicated experience with non-statistical sampling methods.
BIG 4 is measured by the indication of working experience in a Big 4 audit firm.
EXAM_PREP is measured by the indication whether the participant has already prepared for the auditor examination.
KNOW_STAT is measured by self-assessment of individual knowledge towards statistical methods in auditing.

Table 6.5: Test of H1b

Panel A: Full Model[a] ANCOVA (*SAMPLE SIZE*)

Source	SS	df	F	p-value
GUIDANCE	1,820.93	1	3.06	0.0820
TASK RELATED EXPERIENCE	7.58	1	0.01	0.9102
AUDIT EXPERIENCE	1,116.50	1	1.88	0.1725
BIG 4	690.93	1	1.16	0.2826
EXAM_PREP	3,051.38	1	5.13	0.2548
KNOW_STAT	2,464.69	1	4.14	0.0433

Panel B: Descriptive Statistics by Condition

GUIDANCE	Mean	Variance	*Minimum*	Maximum	25%-Quartile	50%-Quartile	75%-Quartile
Present	36.81	618.95	0.0	95.5	15.5	35.5	55.5
Absent	31.32	590.75	0.0	95.5	15.5	25.5	45.5

[a] Between-subjects tests are performed for *GUIDANCE*.
SAMPLE SIZE is measured by the indicated number of sampling items.
GUIDANCE is manipulated between subjects at two levels: Present and absent.
TASK RELATED EXPERIENCE is measured by the indicated experience with non-statistical sampling methods.
AUDIT EXPERIENCE is measured by the indicated years of experience in audit.
BIG 4 is measured by the indication of working experience in a Big 4 audit firm.
EXAM_PREP is measured by the indication whether the participant has already prepared for the auditor examination.
KNOW_STAT is measured by self-assessment of individual knowledge towards statistical methods in auditing.

Panel A of Table 6.4 also shows the results of our H2 and H3. Hypothesis H2 states that reliance on a client's internal controls will influence auditors' decisions on sample size in such a way that recommended sample sizes in high-risk environments will be larger as compared to those in low-risk environments, with the effect being stronger in the presence of *GUIDANCE*. Similarly, hypothesis H3 indicates that recommended sample sizes will be larger when no additional substantive testing procedures toward the same audit objective are performed, with the effect being stronger in the presence of *GUIDANCE*. The main effects of the two variables *RISK* and *PROCEDURES* are highly significant, with *p*-values of 0.0000 and 0.0003 respectively. Thus, both hypotheses H2 and H3 cannot be rejected. However, no significant interaction effects between *GUIDANCE* and either *RISK* or *SAMPLE SIZE* can be observed. This leads us to conclude that the participants are already aware of the influence of *RISK* and *PROCEDURES* on *SAMPLE SIZE*, which is not amplified by the presence of the decision aid.

Finally, we observe a moderately significant interaction effect of *GUIDANCE* with *RISK* and *PROCEDURES* ($p = 0.0846$) in our model (see Table 6.4, Panel A). We investigate this further by performing a simple effects analysis of *GUIDANCE* at different levels of the two other independent variables, *RISK* and *PROCEDURES*. This analysis is shown in Panel B of Table 6.4. The observed effect is driven solely by the condition in which *RISK* is high and additional *PROCEDURES* are absent, with a *p*-value of 0.0210. Regarding the observed interaction, Figure 6.1 shows that at the most acute level of detection risk, the decision aid comprised in ISA 530 prevents auditors from drawing the conclusion to add up the effects of *RISK* and *PROCEDURES*. In other words, the presence of decision aids leads to larger sample size recommendations in any condition except in the most important one, that is when *RISK* is high and *PROCEDURES* are absent.

6.5.4 Controls

Only the *EXAM_PREP* main effect is significant at the $p < 0.1$ level ($p = 0.0971$). As the preparation for the exam serves as a proxy for knowledge and experience, we conclude that a significant effect on *SAMPLE SIZE* can exist only when an auditor possesses in-depth knowledge of sampling procedures. By contrast, the main effects of *AUDIT EXPERIENCE* in years and *TASK-RELATED EXPERIENCE* as a proxy for working knowledge with non-statistical sampling plans are not significant in our model. There are also no significant effects for *KNOW_STAT* or *BIG 4*. This leads us to conclude that any auditor,

regardless of how long they have worked in an auditing environment, their experience with non-statistical sampling methods, or the auditing firm they are employed in, determines sample sizes that are either too large to work economically or too small for one to draw valid conclusions about the underlying population.

Nevertheless, we must interpret these results cautiously, for two reasons. First, our measurements of *EXPERIENCE* and *TASK-RELATED EXPERIENCE* are based only on the number of years working in auditing and on practical experience with non-statistical sampling. Furthermore, neither variable may reflect any actual understanding of task fulfilment or expertise in the matter. Second, *KNOW_STAT* is based on a self-assessed measurement, so we cannot rule out any distortion in the participants' self-evaluation.

6.5.5 Additional Results

Elder et al. (2013) observe that audit firms tend to move from statistical to non-statistical sampling to make it easier to attain smaller samples. Such an action can result in a disregard for statistical demands. We can compare the recommended sample sizes from our experiment to the sample sizes obtained by either (1) a monetary unit sampling (MUS) approach, (2) a quasi-statistical formula offered by the AICPA audit guide, or (3) a tabulated heuristic offered by the AICPA audit guide. The standard setters of all audit sampling standards known to us demand that non-statistical sample sizes be at least comparable to statistical sample sizes. We therefore constructed a hypothetical accounting population, matching the attributes of the experiment population.

For tests of details, sample size is a function of account balance, level of inherent and control risk, assurance from additional tests, tolerable misstatement, expected magnitude and frequency of misstatements and population size. ISA 530 certainly does not include quantitative considerations of these factors. On the contrary, the 2012 AICPA guidance suggests the use of a quasi-statistical formula for sample size determination, relating tolerable error to the account balance, weighted by an assurance factor, while the 2017 AICPA guidance suggests a tabulated sample size determination depending on input factors related to the MUS approach. As a stratification of the population in our experiment had already been performed, we set the risk of incorrect acceptance between 50 percent and 90 percent and confidence factors between 0.7 and 2.7, depending on the level of *RISK* and *PROCEDURES*. For the formula-based approach, we accepted a sample size penalty at the 0.25 level, as a non-statistical sampling approach was used and error

expectation was greater than zero. As all other factors were held constant in our case, we compare computed sample sizes to those from the experiment shown in Figure 6.1.

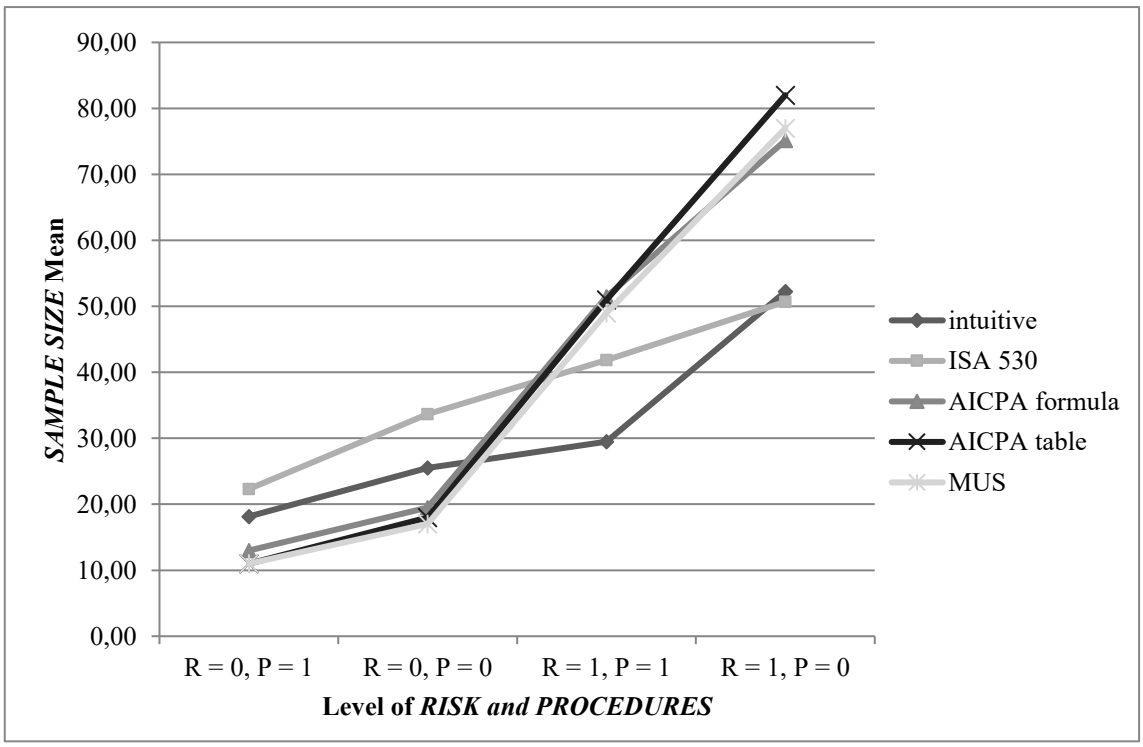


Figure 6.1: *SAMPLE SIZE* Comparison by Procedure

Two key observations can be made here. First, sample size means for the four scenarios of *RISK* and *PROCEDURES* are higher for statistical methods than for non-statistical methods. In other words, participants using non-statistical approaches evidently recommend smaller sample sizes, when in fact they should draw larger samples because of the penalty already mentioned. As Figure 6.2 shows, this holds for both levels of *GUIDANCE*. Second, sample size variances between conditions are larger for statistical methods than for non-statistical methods, which indicates a higher sensitivity to changes in the level of risk and to the presence of further substantive testing procedures, which strongly supports the anchoring and adjustment behavior.

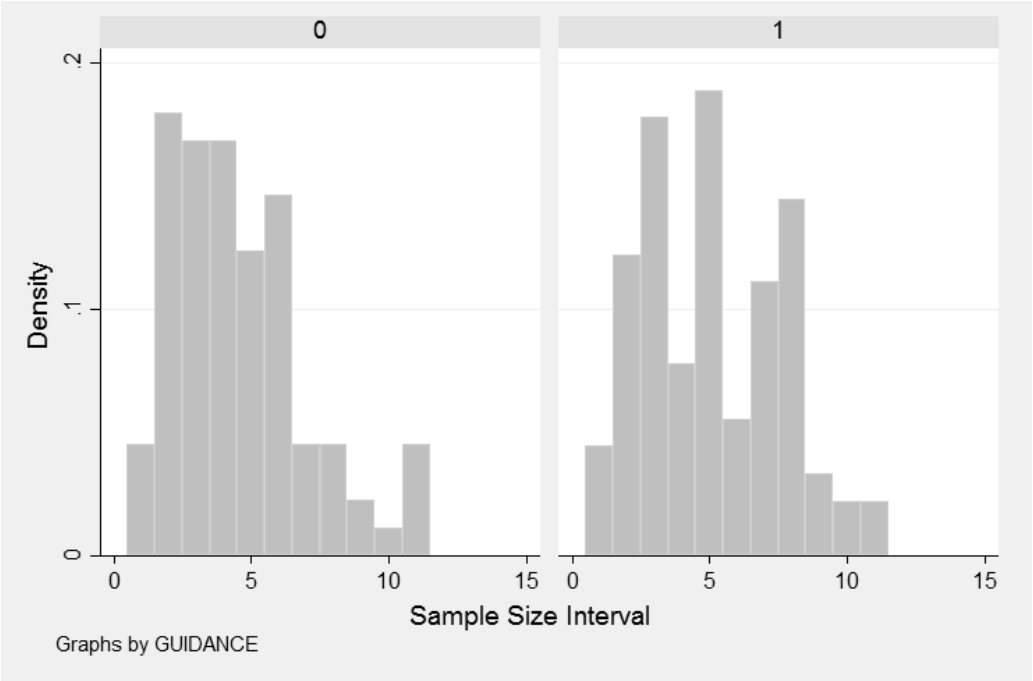


Figure 6.2: Density function of *SAMPLE SIZE* by condition of *GUIDANCE*

6.5.6 Discussion

The results of our empirical analyses show that the decision aid provided by ISA 530 significantly impacts participants' judgments about the magnitude of sample sizes but does not improve auditor consensus on accurate sample sizes. Our results hold for two levels of client risk as well as for two levels of the extent of additional procedures performed towards the same audit objective. These findings support our argument that decision biases can be mitigated when a participant's cognitive strain is moderated.

However, absolute differences between the tested conditions are relatively small. In line with previous literature, our findings indicate that auditors are to some degree already intuitively aware of the factors considered in ISA 530 (Kachelmeier and Messier 1990; Messier et al. 2001; Durney et al. 2014; Christensen et al. 2015). Compared to statistical methods, the decision aid intensifies the danger of being overly optimistic in low-risk environments by performing only few, if any, substantive tests. The lack of quantitative measures in ISA 530 seems to lead auditors to underestimate the extent of testing that is necessary for one to draw conclusions that at least approximate statistical concepts. However, we doubt that auditors on the staff or senior level are aware of the need to arrive at sample sizes comparable to statistical theory. While Messier et al. (2001) found a need to improve the AICPA audit guide formula, from our perspective the ISA guidance should be complemented by additional support that specifies the qualitative requirements of ISA

530 in a quantitative way. These needs are evident, considering that non-statistical sampling suppresses statistical approaches in the field of auditing.

The results of our experiment confirm that anchoring behavior in the judgmental determination of sample sizes is prevalent. We provide the following explanation of auditors not being adequately sensitive to varying input factors. An internal anchor primarily constitutes an auditor's consistent 'rule of thumb', as most clients have fair to strong internal control over financial statements. Additionally, auditors usually perform several audit procedures to achieve the same auditing objective. When auditors calculate sample sizes based on average-experience parameters, an a-priori expectation is formed. Non-familiarity with divergent auditing situations leads to insufficient adjustments, as the normatively expected adjustment surpasses the subjective consideration. Messier (1995) predicts that without decision aids auditors' approaches to addressing a sophisticated problem can vary largely. Hence, unaided judgment should be made only when an available decision aid is not employed (Ashton and Willingham 1988). In the case of non-statistical sample size judgments, aided judgment is a reasonable option, as auditors seem to be insensitive to certain base rate information and lack decision consensus.

6.6 Conclusions, Limitations and Future Research

We examine one of the most frequently discussed concerns in auditing practice: that audit sampling in the field regularly does not have the validity it is attributed with, primarily because of insufficient sample sizes. Our study therefore investigates the potential effect of a semi-structured decision aid on auditors' sample size determinations. We find that sample sizes are significantly larger under ISA 530 guidance than by intuitive judgment, but remain insufficient when compared to statistical sampling procedures. The standard's guidance is quite scant, which demonstrates the need for additional instruction from the IAASB. We agree with Sibelman (2014), who says that even under AICPA regulation more guidance is needed. Our results indicate that unaided auditors tend to agree on initial sample sizes in non-statistical audit sampling applications, employing 'rules of thumb' that are not sufficiently adjusted to client-specific properties. The results hold for two levels of inherent and control risk (high vs. low) as well as for two levels of the extent of additional testing procedures (present vs. absent), implying that the disclosed uniform strategy favorably leads to under-auditing in high-risk environments. When auditors neglect to gather reliable evidence that is both appropriate and sufficient to express the

predicated audit opinion, they are subject to regulatory sanctions as well as civil or criminal penalties imposed by trial courts.

Our experiment provides important empirical contributions to the theoretical discussion of the presented hypothesis. The study therefore contributes to the growing body of literature on auditor behavior by extending research on auditors' performance in substantive testing, analyzing the effect of changing auditing standards in terms of an effective move towards a greater judgmental consensus. Moreover, we shed light on the assumption that auditors tend to overly rely on personal experience, being insensitive to unfamiliar audit situations. Hence, our results offer important practical implications for auditing firms and regulators. Firms should be aware of the constraints revealed and counteract them by implementing meaningful internal guidance going beyond normative statements. Moreover, regulators can benefit from the presented findings, as they might consider an improvement in the current wording of auditing standards.

We emphasize that our investigation is subject to some limitations associated with the experimental design and procedure. Apart from a possible lack of task realism, participants were not selected randomly. The participants' work has not been reviewed by any superior member of an audit team, hence non-fulfillment of requirements could not be sanctioned, or audit procedures revised. Results may also be limited because of a possible imprecision in the measurement of control variables, as task-related experience and statistical knowledge were assessed by the participants themselves, resulting from time and anonymity constraints.

Overall, we believe that the investigation of decision aids in non-statistical sampling plans in terms of sample size determinations enriches the existing behavioral literature. We encourage future research into auditors' decision making in substantive testing and the tendency to bypass normative guidance. Future research might also examine potential debiasing techniques and test alternative decision aids directed toward mitigating these concerns.

6.7 Appendix: Instrument

Note: Each subject received the following case. Wording in parentheses comprises information included in the low-risk condition. Throughout the introductory narrative,

- wording in square brackets comprises information included in the high-risk condition,
- italicized wording comprises information in the additional audit procedures condition,
- bold, italicized wording comprises information in the condition without additional audit procedures performed.

The table comprising factors considered by ISA 530 was provided only for subjects in the ISA 530-guided condition (including any wording in parenthesis or square brackets).

Welcome!

Thank you for participating in our research study on the efficacy of audit regulation in the environment of audits of financial statements. For the purpose of our research, we will survey individuals performing audits of financial statements and those who already focus their tenure throughout their study courses.

It is crucial for our research results to engage a sufficient number of participants with professional practice. Therefore, we will donate an amount of 5 € for every completed survey to a charitable institution of your choice, which can be chosen at the end of this survey.

You will perform the case study from an auditor's perspective. Succeeding the introduction, you will receive general information about the annual financial statements for the fiscal year 2016 of the fictional company "Cookie Products Ltd.". Subsequently, you will obtain a narrative of a specific issue of an audit procedure, which will require your assessment. We ask you to take the perspective of the auditor of "Cookie Products Ltd." throughout the case study. The time needed to read all necessary information and to finalize the case study will be about 15 to 20 minutes.

Please reply honestly to all questions in the survey. There are no "right" or "wrong" answers. We are simply interested in observing your true behavior. All answers will be kept confidential; the processing of the case study is completely anonymous.

We gratefully acknowledge your support!

page turn

COOKIE PRODUCTS LTD.

Assume you are the auditor in charge for the audit of the annual financial statements of Cookie Products Ltd., a manufacturer of kitchen appliances. The unadjusted numbers of the fiscal year 2016 show the following balance sheet and income statement:

Cookie Products Ltd.
Balance sheet as of 31st of December 2016
in T€

Assets		**Liabilities & Stockholders' Equity**	
Non-current assets:			
Technical equipment and machinery	7.072	Stockholders' equity:	
		Capital stock	8.032
Other equipment, factory and office equipment	5.056	Retained earnings	13.958
Total non-current assets:	**12.128**	**Total stockholders' equity:**	**21.990**
Current assets:			
Inventory of manufactured goods	5.264		
Inventory of supplies	5.231		
Accounts receivable	***6.758***	Liabilities:	
		Long-term debt	5.222
Cash	3.402		
		Current liabilities	5.571
Total current assets:	**20.655**	**Total liabilities:**	**10.793**
Total assets:	**32.783**	**Total liabilities & stockholders' equity:**	**32.783**

Cookie Products Ltd.
Income statement as of 1st January 2016 – 31st December 2016
in T€

Sales	25.494
Cost of materials	13.041
Gross profit	12.453
Personnel expense	2.300
Selling & administrative expenses	5.032
Net income before taxes	5.121
Provision for taxes	2.356
Net income	2.765

Net income has been stable in the range of 2 to 4 Mio. € for the past 5 years.

page turn

Your audit attention is directed towards the balance of accounts receivable of T€ 6.758. This amount represents 978 debtors, whereas individual debtor balances vary in their amount between T€ 1 and T€ 749.

Existence, accuracy, and valuation of accounts receivable shall be audited by confirmations of balances, which shall be obtained from a sample. For audit purposes, the amount of accounts receivable is considered material to the financial statements taken as a whole but is not so critical as to warrant formal statistical analysis. Determined deviations, which are caused by incorrect booking of the client, shall be projected to the population, the audit objective "accounts receivable".

Complementing a judgmental selection of debtors with high amounts of balances, the audit manager determines the application of non-statistical sampling with random selection of the remaining population to be sufficient:

- Complete-count stratum: A judgmental selection of debtors with 6 highest sole balances of altogether T€ 2.027 achieving a coverage of 30 % of the entire balance of accounts receivable.
- Selective stratum: The remaining population consists of 972 debtors with a jointly balance of T€ 4.731. Individual debtor balances vary between T€ 1 and T€ 50. A random selection of *n* additional debtors shall be performed.

The sample size in the selective stratum shall be kept preferably small due to the expected quantity of alternative audit procedures. Nevertheless, sufficient audit evidence shall be obtained in any case to form an opinion about the balance of accounts receivable. On the basis of the audit approach of your audit firm, the guideline of the responsible audit partner is not to collect more than a maximum of additional 100 confirmations of balances.

Hereafter, the case study demands from you to consider various characteristics of the audit field of accounts receivable, which are significant to the elicitation and evaluation of an audit sample. In so doing, please consider the information given on the next slide.

page turn

Materiality

Overall materiality threshold is T€ 510 or 2 % of sales revenue.

Internal control

The external accounting department of Cookie is staffed by (13) [5] employees, thereof (3) [1] accounts receivable accountants. All accounting records and systems are fully electronic in a unified ERP-system, whereby transactions concerning debtors are automatically adopted into the general ledger.

Audit of the internal control system has (not) led to complaints. A sample of 30 accounting transactions within tests of internal control yielded to (no) [4] recognized errors. [All detected errors occurred because orders were accepted although there wasn´t any creditworthiness of the customers and because of defective processing of credit notes.]

An examination of the accounting journal disclosed that – apart from booking global and specific valuation allowances – (only) [besides] the authorized accounts receivable accountant(s) [also the head of accounting] made booking entries within accounts receivable accounts.

New control processes

Due to (immaterial) [material] control deviations that were revealed in the preceding years audit, two new control processes were implemented (at the beginning of the financial year) [during the financial year] as a reaction to recommendations from the prior-year management letter.

First, a documented authorization concept for the maintenance of master data was established. All master data including lines of credit and banking information are still maintained by the authorized accounts receivable accountants; however, any changes made to those data are now logged and need to be additionally approved by a supervisor.

Second, quarterly aging analyses of accounts receivable are now performed, to ensure that payment reminders are promptly sent to debtors and irrecoverable accounts receivable are identified early.

Tests of these new controls indicate that control procedures appear to be operating as described throughout the whole audit year. [Tests of these new controls indicate that control procedures were implemented during the year and therefore were not effective throughout the whole audit year. Hence, functional testing of newly implemented controls did not occur.]

Nature of Other Substantive Tests

Besides requesting confirmations of balances, further audit procedures are scheduled to form an audit opinion about the accounts receivable. For one, analytical audit procedures (trend analysis, validation of plausibility) were performed to support the accounted balance of accounts receivable. Additionally, audited were revenues concerning recognition and existence during the year and concerning year-end cut-off procedures near the balance sheet date. ***Besides requesting confirmations of balances, no further audit procedures are scheduled to form an audit opinion about the accounted balance of accounts receivable.***

Prior-year audit conclusions

Results from last year's similar audit test of accounts receivable showed a moderate number of misstatements, usually overstatements, which amounted to 1 to 1.5 % of the account balance.

page turn

Please note the following factors and their direction of action toward the determination of the necessary sample size for tests of details:

Factor	Effect on Sample Size	
1. **An increase in the auditor's assessment of the risk** of material misstatement	Increase	The higher the auditor's assessment of the risk of material misstatement, **the larger the sample size needs to be**. The auditor's assessment of the risk of material misstatement is **affected by inherent risk and control risk**. For example, if the auditor does not perform tests of controls, the auditor's risk assessment cannot be reduced for the effective operation of internal controls with respect to the particular assertion. Therefore, in order to reduce audit risk to an acceptably low level, the auditor needs a low detection risk and will rely more on substantive procedures. The more audit evidence that is obtained from tests of details (that is, the lower the detection risk), the larger the sample size will need to be.
2. An increase in the use of **other substantive procedures** directed at the same assertion	Decrease	The more the auditor is relying on other substantive procedures (**tests of details or substantive analytical procedures**) to reduce to an acceptable level the detection risk regarding a particular population, the less assurance the auditor will require from sampling and, therefore, **the smaller the sample size can be**.
3. An increase in the auditor's **desired level of assurance** that tolerable misstatement is not exceeded by actual misstatement in the population	Increase	The greater the level of assurance that the auditor requires that the results of the sample are in fact indicative of the actual amount of misstatement in the population, **the larger the sample size needs to be**.
4. **An increase in tolerable misstatement**	Increase	The lower the tolerable misstatement, **the larger the sample size needs to be**.
5. An increase in the **amount of misstatement the auditor expects** to find in the population	Increase	The greater the amount of misstatement the auditor expects to find in the population, **the larger the sample size needs to** be in order to make a reasonable estimate of the actual amount of misstatement in the population. **Factors relevant** to the auditor's consideration of the expected misstatement amount include • **the extent to which item values are determined subjectively** • **results of risk assessment procedures** • **results of tests of control** • **results of audit procedures applied in prior periods** • **results of other substantive procedures.**
6. **Stratification of the population** when appropriate	Decrease	When there is a wide range (variability) in the monetary size of items in the population, it may be useful to stratify the population. When a population can be appropriately stratified, the **aggregate of the sample sizes from the strata generally will be less** than the sample size that would have been required to attain a given level of sampling risk, had one sample been drawn from the whole population.

7. **The number of sampling units** in the population	Negligible effect	For large populations, the actual size of the population **has little, if any, effect on sample size**. Thus, for small populations, audit sampling is often not as efficient as alternative means of obtaining sufficient appropriate audit evidence. (However, when using monetary unit sampling, an increase in the monetary value of the population increases sample size, unless this is offset by a proportional increase in materiality for the financial statements as a whole [and, if applicable, materiality level or levels for particular classes of transactions, account balances or disclosures].)

page turn

Objective

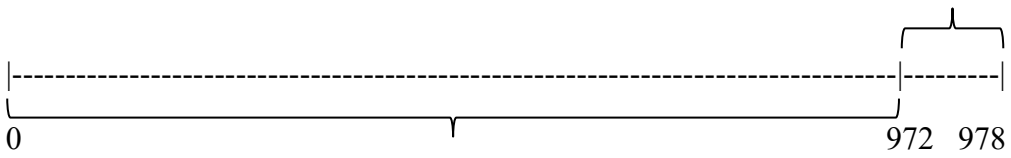
Please indicate your intuitive judgment of the desired number of sample items for the detailed test of the accounts receivable balance just described, corresponding the number of confirmations of balances to be requested.

Please remember to consider the initial descriptions of balance sheet and income statement values, audit conclusions regarding the internal control system, the information given to other substantive audit procedures, and the expectation of misstatements due to prior-year audit conclusions.

The possible amount of confirmations of balances to be requested is determined by the number of debtors representing the balance of accounts receivable, whereas the maximum limit due to the company´s audit approach and instruction by the lead audit partner is 100:

Complete-count stratum

6 debtors with highest individual balances



Remaining stratum for audit sampling (random selection)

Within the scope of the complete-count stratum, 6 confirmations of balances are requested. In your opinion, how many additional confirmations of balances should be requested from the remaining stratum to be sufficient audit evidence?

(0) 0
(0) 1 to 10
(0) 11 to 20
(0) 21 to 30
(0) 31 to 40
(0) 41 to 50
(0) 51 to 60
(0) 61 to 70
(0) 71 to 80
(0) 81 to 90
(0) 91 to 100

page turn

Please briefly indicate the reasoning you used to arrive at this judgment. Why didn´t you choose to request more or less confirmations of balances?

page turn

Further Questions

Please answer the following questions faithfully and honestly:

1. As far as you remember: Was a part of the accounts receivable balance of Cookie Products Ltd. entirely tested within the scope of a complete-count-stratum?

1 (I don´t agree)	2	3	4	5	6	7 (I fully agree)
O	O	O	O	O	O	O

2. As far as you remember: How many additional substantive tests of details with the same audit objective as the confirmations of balances were performed during the audit of Cookie Products Ltd.?

1 (none)	2	3	4	5	6	7 (many)
O	O	O	O	O	O	O

3. A deficient internal control system leads to a higher probability of material misstatements in financial reporting. In your opinion, how reliable is Cookie Products Ltd.´s implemented control structure in the former case?

1 (not reliable at all)	2	3	4	5	6	7 (very reliable)
O	O	O	O	O	O	O

4. As far as you remember: Which normative advices to determine an accurate sample size did you receive additionally with the general information about Cookie Products Ltd.?

   (O) I did not get any more advice for the case study
   (O) I got additional advice from ISA 530 regarding factors influencing the sample size for substantive tests of details

5. [ONLY IF 4 = YES] How much did the additional advice from ISA 530 help you determine an accurate sample size?

   (O) Helped me a lot
   (O) Helped me a little
   (O) Did not help me

6. How precisely did you read the case study of Cookie Products Ltd. and the complementary information and premises to the case study (Please reply honestly. None of the following answers is “right” or “wrong“; we are simply interested in your true behavior.)?

(O) not read / bypassed
(O) briefly / read across
(O) fairly intensive (word for word / number for number)
(O) very intensive (e. g. repeatedly, tried to memorize content)

7. Did you already have practical experience with comparable or similar audit procedures?

1 (I don´t agree)	2	3	4	5	6	7 (I fully agree)
O	O	O	O	O	O	O

8. For how many years have you already worked in the audit field?

_____ years

9. In what kind of audit firm are you employed or you were employed in?

(O) Big 4
(O) Non Big 4

10. In the audit of what kind of enterprises do you have field experience?

(O) private companies / partnerships
(O) small and medium-sized enterprises
(O) large corporations pursuant to Sec. 267 (3) HGB
(O) Public Interest Entities (e. g. capital market listed enterprises, banks, insurances etc.)

11. Did you already prepare the “Wirtschaftsprüfer” exam?

(O) Yes
(O) No

12. How frequently do you have to deal with sampling in the context of auditing?

1 (never)	2	3	4	5	6	7 (often)
O	O	O	O	O	O	O

13. How frequently do you use formal statistical sampling methods, e. g. Monetary Unit Sampling, in the field?

1 (never)	2	3	4	5	6	7 (often)
O	O	O	O	O	O	O

14. How frequently do you use non-statistical sampling methods, e.g. random selection from fixed sample sizes that are not deduced by mathematical formulas in the field?

1 (never)	2	3	4	5	6	7 (often)
O	O	O	O	O	O	O

15. How frequently do you use the selection of specific elements by professional judgment, e.g. highest individual balances, in the field?

1 (never)	2	3	4	5	6	7 (often)
O	O	O	O	O	O	O

16. How would you rate your skills in mathematics and statistics?

(O) no skills, only hands-on experience
= application possible, comprehension of statistical coherences vague
(O) moderate skills
= essential statistical knowledge available, application mandatory
(O) extensive skills
= comprehension of methods and application mandatory

7 How Many Needles Are in the Haystack? Sensitivity of Error Containment Judgments to Changes in Audit Standards

7.1 Publication Details

Abstract: Professional standards in audit sampling require the auditor to project errors to the population sampling items are drawn from. However, non-projection of sampling errors is everyday practice among auditors in the field, increasing overall audit risk and the possibility of issuing an inappropriate audit opinion (Christensen et al. 2015). This paper examines the effect of error containment guidance on auditors' containment judgments and the perceived occurrence of anomalies in audit sampling. Referring to the representative heuristic, we hypothesize that general as well as explicit error-specific guidance reduces auditors' tendency to remove misstatements from an estimate for total error, even when initially provided with the potential presence of an error. We test our hypothesis with an experiment among 80 students. Contrary to expectation, we find that additional guidance does not significantly improve bias avoidance. Our results also indicate that containment decisions are primarily influenced by error-specific properties such as error frequency and the amount of containment information available, rather than the actual cause of an error. Hence, even in common audit settings, we find participants lacking appropriate containment decisions.

Co-Authors: Prof. Dr. Christiane Pott.

Keywords: audit sampling, error containment, representative heuristic, audit misstatements, auditor judgment.

JEL-Code: C91, M42, D91.

Publication Status: Working paper.

7.2 Introduction

In this paper, we investigate the effect of error-specific properties and their consideration in audit sampling guidance on auditors' error containment judgments. With non-statistical audit sampling being the most common practice in the field (Durney et al. 2014), the evaluation of revealed sampling errors is much more prone to decision biases (Christensen et al. 2015; Elder et al. 2013), that is, compared with statistical sampling plans, owing to the lack of a statistical inference on errors and the omission of mathematical sampling risk estimations. Although the isolation of sampling errors is a known concern in auditing practice (Christensen et al. 2015), ISA 530 is the only auditing standard which provides explicit guidance on error containment. The standard will become mandatory in German audits from 2018, raising the question of whether more guidance on this topic leads to changes in auditors' decisions to contain material misstatements revealed by audit sampling procedures. Determining the effects of enhanced guidance on auditor behavior is important, as previous research has shown that the isolation of errors remains a major concern in the field, and that changing auditing standards do not necessarily increase audit quality (Christensen et al. 2015; Durney et al. 2014; PCAOB 2008; Messier et al. 2001; Burgstahler et al. 2000). It is important to understand how auditors evaluate and resolve sampling errors, with the projection of representative misstatements being one of the most important decisions within a sampling plan (AICPA 2017; Durney et al. 2014). A better understanding of current error containment practice and its sensitivity to changing audit standards is overdue, as virtually any financial statement audit involves sampling to gather substantive audit evidence (Krahel and Titera 2015; Sibelman 2014). Consequently, current auditing standards require the auditor to provide a reasonable basis on which to draw conclusions about the entire population (PCAOB 2014; AICPA 2011; IAASB 2009b).[178] As the evaluation of aggregate error is highly dependent on the projection of generally any revealed errors, this highlights the potential hazard of arbitrary error containment.

Previous research has shown that auditors have difficulty in applying the statistical principle of extrapolation and circumvent this by arguing that an error can be identified (1) as unique and therefore not representative of the underlying population, and (2) as occurring only in a segment of the population, with the assumed segment being then investigated in

[178] Conversely, if a subset of items is not representative of a population, a sample does not provide more than a description of a part of that population.

more detail (Burgstahler and Jiambalvo 1986; Dusenbury et al. 1994; Burgstahler et al. 2000). Both actions lead to a non-projection decision about revealed errors. This exercise can increase the risk of a materially misstated population being accepted, when the error is in fact representative of the underlying population. In contrast to the former normative status in Germany, ISA 530 now includes qualitative guidance on the evaluation and resolution of audit samples, explicitly considering the possibility of errors being labeled as anomalies.[179] Relevant to the current study is the fact that the transformation into local GAAS will potentially make justifying non-projection decisions easier (Burgstahler et al. 2000), leading to less accurate error projections.

For two types of error (common errors and fraudulent errors), we examine the effect of the presence of the ISA guidance and the impact of a key component of this guidance: containment information, which is acquired by performing additional audit procedures on a subset of the initial sample drawn from an audit population. Containment information might be perceived as a reasonable conclusion in containing errors, especially when an uncommon reason for the errors is revealed. We test our research question using a 2 x 2 x 2 mixed between- and within-subjects experiment, in which we manipulate the extent of ISA 530 guidance (absence vs. presence) as well as two levels of containment information (absence vs. presence) and two levels of error frequency (high vs. low). A total of 80 senior-level accounting students assessed the necessity of error projection for eight different errors revealed within the scope of a prototypical non-statistical sampling plan. We find that, even when the guidance is present, participants significantly underestimate the implications of projected errors and contain misstatements that cannot be classified as anomalies. Results also indicate that perceived error frequency and the amount of containment information directly affect the probability of containing an error and overriding the effect of the presence of guidance. In line with prior research in psychology and auditing (Christensen et al. 2015; Durney et al. 2014; Elder et al. 2013; Allen and Elder 2005; Hitzig 2001; Burgstahler et al. 2000; Elder and Allen 1998; Hermanson 1997; Dusenbury et al. 1994; Kachelmeier and Messier 1990; Burgstahler and Jiambalvo 1986; Butler 1986; Kahneman and Tversky 1972), we conclude that, owing to the favored use of non-statistical sampling, decision biases and heuristics are more prevalent than ever when it comes to the evaluation and resolution of errors in audit sampling, and that the

[179] Anomalies are referred to as misstatements or deviations that are demonstrably not representative of misstatements or deviations in a population (ISA 530.5 e).

potential guidance aimed at mitigating possible negative effects must be much more specific than that comprised in ISA 530.

This paper contributes to the auditing literature in several ways. First, we examine whether additional guidance on error evaluation and error containment lowers an individual's tendency to contain representative errors, potentially leading to higher audit quality. Second, we combine this track with an investigation of the effect from a recently changed audit standard. We test our hypotheses for different types of error as well as two levels of containment information, to estimate the effectiveness of the auditing standard change in various practically relevant audit situations. To the best of our knowledge, no study has tested the implication of ISA 530 guidance on auditor behavior in consideration of error containment.

Third, our research indicating potential flaws in the performance of sampling plans is important for researchers, educators, regulators and practitioners. We provide a better understanding of auditor performance in non-statistical sampling, which today is the preferred sampling method. As the documentation of judgmental sampling decisions is consistently being reviewed by internal and external reviewers, the examination of containment decisions is an important practical issue, as flawed sampling plans can lead to cases of legal liability. Regarding current auditing standards, our findings may inform regulators of necessary improvements. They may also be useful for audit practitioners and indicate starting points for the development of decision aids, as such aids can help auditors deconstruct complex cognitive challenges such as the resolution of sample errors. Finally, as the handling of errors must evolve with the understanding of what typical accounting errors are, our findings highlight areas for future research and emphasize inevitable matters in auditing education.

The remainder of this paper is organized as follows. In the next section, sample error evaluation and containment in the auditing context are discussed, we review prior literature on this topic, and describe the development of our hypothesis. The third section presents our research design and data collection. Results and implications are discussed in the fourth section. The final section concludes the paper by summarizing the results and discussing study implications.

7.3 Background and Hypothesis Development

7.3.1 Error Evaluation and Containment in Audit Sampling

As audit sampling in general comprises the use of statistical or non-statistical sampling techniques, the latter are common in the field (Christensen et al. 2015; Elder et al. 2013). Non-statistical approaches follow either a non-random selection of sample items or a non-probabilistic evaluation of sample results. Irrespective of which approach is used, an accurate conclusion of sampling results must include the projection of revealed misstatements to the underlying population (ISA 530.14). While in attribute samples, such as tests of internal control over financial statements, the sample deviation rate equals the expected population deviation rate, error projection is always relevant when testing account balances. Hence, for substantive testing, PCAOB audit standards as well as the AICPA audit sampling guide indicate the necessity of error projection in any case (AU-C 530.13; AICPA 2017), whereas ISA 530 indicates that in "extremely rare circumstances" errors might be treated as anomalies and hence do not need to be projected, if the auditor is able to obtain evidence that a misstatement is not representative of the population (ISA 530.13, A19, A22). However, error containment is observed to be common practice, even beyond ISA-regulated audit settings (Baumeister et al. 2018; Christensen et al. 2015; Fay et al. 2015; Durney et al. 2014).[180]

The evaluation of a sample error[181] can be a sophisticated task, with various and partially counteracting parameters to be considered, e.g. the cause of the error, the control environment surrounding the cause,[182] as well as additional audit work that has been performed. To mitigate the difficulty of the arising exercise auditors might tend to reduce this complexity by relying on personal experience and a limited number of heuristic principles that simplify certain judgmental operations, but they will also increase the risk of severe errors. If an auditor discovers a sample error, it is likely that additional errors exist in the remaining items of the account balance. Thus, revealed errors proxy for additional

[180] The AICPA audit guide determines that "an auditor plans and evaluates an audit sample with the knowledge that the overall conclusion about the population characteristic of interest is based on more than the results of that audit sample" (AICPA 2017, 2.06). This phrase might be interpreted as implicitly permitting error containment.

[181] As in Burgstahler and Jiambalvo (1986), the word "error" in our study comprises both the misstatement found in an accounting population and its cause. Similarity under these circumstances means that the cause of a revealed error is similar to other potential causes of errors in the same population, and therefore has large causal similarity (Tversky 1977).

[182] For example, the possibility of errors made on purpose, and the presence of internal control mitigating such scenarios.

errors. Previous empirical studies have provided evidence that the treatment of sample errors is worrying when it comes to error projection decisions (Burgstahler and Jiambalvo 1986; Elder and Allen 1998; Burgstahler et al. 2000). As a result, misconceptions in the evaluation of sampling results eventuate in an incorrect acceptance of balances (Allen and Elder 2005). In particular, individually immaterial errors can become material when projected onto the total population. ISA 530.12 demands that "the auditor shall investigate the nature and cause of any deviations or misstatements identified and evaluate their possible effect on the purpose of the audit procedure and on other areas of the audit", driving auditors to treat qualitatively different errors in quantitatively different ways.

To examine this issue more precisely, we extend a normative model for the arbitrary exclusion of errors, by taking a balls and urns example, based on Burgstahler and Jiambalvo (1986). Of n items drawn from a population of N balls, if r balls are red and $(n - r - 1)$ balls are black, and one ball is red, but rather shaped like a cube than a sphere, the possibility of drawing that cube from the population is small and becomes more unlikely with an increase in N. Hence, auditors may expect that a revealed error assumed to be unique (red cube) is not to be found elsewhere in the population. What the auditor misses is that the red cube proxies for other red items, even of different shapes (e.g., a cone) comprised in the population. Just because these red items vary in shape (cause of the error), this does not mean they should be treated differently when it comes to the projection decision. Consequently, the example would infer that all errors could be isolated, making the intent of audit sampling useless. Owing to various possible qualitative aspects of errors in practice (causes of errors), any error in auditing is somewhat unique. As containment scenarios are particularly prevalent when auditors perform discovery sampling,[183] an auditor might claim that a population does not contain any error at all, if no errors are found in a sample. Conversely, when a single error is revealed, the auditor might tend to call this error the only one in the population. While errors in accounting populations are effectively rare (Durney et al. 2014), it is easy to understand that small samples are not very effective in disclosing those rare items. In a population of 500 items with two items being misstated, comprising a deviation rate of 0.4%, the odds against finding one of these

[183] Selecting small samples while expecting no errors.

misstatements by drawing a sample of 30 items is $\frac{500-30}{30}/2 = 8$ to 1.[184] Drawing such conclusions is risky without performing additional audit work.

From a psychological view, projection decisions are affected by similarity perceptions,[185] for which the representative heuristic provides the adequate basis (Kinney and Uecker 1982; Uecker and Kinney 1977; Tversky 1977; Tversky and Kahneman 1974). The theory is found to be prevalent in several cognitive biases linked to sampling procedures. Previous research in auditing (Burgstahler and Jiambalvo 1986; Butler 1986; Hitzig 1995; Kachelmeier and Messier 1990; Dusenbury et al. 1994; Elder and Allen 1998; Hermanson 1997) and psychology (e.g. Kahneman and Tversky 1972; Tversky and Kahneman 1971; Glass et al. 1979) shows that auditors regularly make mistakes in evaluating samples, as their judgments are influenced by certain sample properties and error characteristics. Relevant to error projection decisions, two cognitive biases within the representativeness heuristic lead to a prototype matching approach, resulting in the use of categories (Nisbett and Ross 1980; Glass et al. 1979). The two biases are the following:

1. Decisions are influenced by salient or concrete data (Kahneman and Tversky 1972). An error found in a sample is more concrete than the hypothesized value of an error projected to the population, leading one to give insufficient attention to projected (aggregated) error.

2. The belief in the "law of small numbers" (Tversky and Kahneman 1971; Kahneman and Tversky 1972) makes individuals assume that samples drawn from a population will be more akin to that population than statistical theory predicts. If an individual's confidence in the representativeness of a sample is arbitrarily exaggerated, the intrinsic uncertainty, when he or she is forming conclusions on a sample basis, will be underestimated (Kachelmeier and Messier 1990).

[184] In addition to a formula provided by Hitzig (2001), the odds against detection of an isolated error generally are $\frac{(N-n)}{n}/z$, with N being the number of items in the population, n being the sample size and z being the actual number of errors in the population. Complementing the example, if the population consists of 2,500 items and the number of actual errors as well as the sample size are held constant, the odds against detection of at least one error are $\frac{2500-30}{30}/2 = 41$ to 1, and so on. As accounting populations usually consist of large numbers of individual items, the probability of finding a rare but possibly material error is very small, and therefore also the probability of reasonable containment.

[185] In the evaluation of sample errors, an auditor asks whether "object A belongs to class B", which implies the further question, "Are the attributes of A representative of attributes of class B". When the probability that an error originates from a class is judged to be low, the auditor assumes that it is not similar to the class.

As causes of accounting errors change, subject to technological advancement and changes in regulation, individuals have problems with error evaluation, especially when certain errors are less similar to other errors they usually expect to find in a similar auditing context.[186]

7.3.2 Literature Review

Major normative and technological changes make it hard to identify specific problem areas to be addressed by auditing firms and regulators.[187] However, a number of studies have examined changes in the quantity and quality of errors occurring in accounting populations, as well as a deficiency in auditors' performance when considering the effects of these changes on their auditing approach.

It is generally accepted that revealed errors proxy for further errors in a population, hence it is usually required to project errors. However, Burgstahler and Jiambalvo (1986) found that 67% of practicing auditors fail to project errors that are actually representative of a population. Akresh and Tatum (1988) and Hitzig (1995) recognized that the tendency not to project errors intensifies when non-statistical sampling plans are used, raising concerns from a present-day perspective, as those methods are the standard nowadays.

Dusenbury et al. (1994) and Wheeler et al. (1997) complemented the research of Burgstahler and Jiambalvo (1986) and found that the probability of projecting an error increases with conventionality, and is less likely when an error is perceived to be containable to a subpopulation, comprising elements with similar characteristics. Their results indicate that auditors have a common strategy in doing so, by (1) identifying a subpopulation to which a sample error can be contained, and (2) investigating (only) this subpopulation for additional errors with the same cause of error. In this way, auditors utilize estimators biased towards error frequency.

Certainly, a biased estimator can be more precise than an unbiased one, initially legitimizing error containment, but proving its validity is complicated. By the time the latter

[186] Changes in the nature and causes of accounting errors are highly topical since the passage of SOX and increased institutional oversight by either the AICPA or PCAOB (Durney et al. 2014). Currently, material errors tend to occur increasingly within internal controls and are more process-driven. However, apart from public interest entities, we expect errors in book entries to remain a key element of any audit.

[187] For a comprehensive review of the related literature focusing on sampling method development, see Elder et al. (2013) and Durney et al. (2014). In line with these studies we assume the related literature to be outdated, as developed estimators were nearly unexceptional based on distributions deviating from actual accounting populations and none have found their way into audit standards or practice.

studies were conducted, any effective auditing standard required error projection. Thus, there was no rationale for why auditors considered such action to be appropriate. Burgstahler et al. (2000) renewed the revealed concerns and attributed them to a lack of auditors' understanding of statistical concepts.

Considering planned materiality as a crucial input parameter for the application of sampling plans, Elder and Allen (1998) reviewed concrete audit working papers and found auditors to contain errors more often if an individual error was large. Hence, error isolation is regularly not used in order to completely avoid error projection per se, but to stratify the population ex post. Focusing on the same audit engagements in subsequent fiscal years, a follow-up study by Allen and Elder (2005) revealed that error projection rates diminished even more, owing probably to increased competition in the audit market. Consequently, PCAOB inspections focused on non-statistical audit sampling applications, confirming the findings from prior experimental research (PCAOB 2008, 2014). In contrast to the PCAOB findings, Durney et al. (2014) found that error projection rates improved in post-SOX applications.

Most recently, in an ongoing survey of auditing firms' policy departments, Christensen et al. (2015) examined sampling guidelines. Sampling experts' internal reviews reported that error projection and the resolution of identified misstatements were two of the biggest difficulties for auditors in the field. One firm found that the frequency of circumventing error projection comprised nearly half of the time sampling was applied in substantive testing, and that often no consultation outside the engagement team was requested.

Supporting these findings, an archival study by Baumeister et al. (2018) shows that average sample sizes of statistical and non-statistical procedures are quite similar. However, sample sizes achieved for non-statistical sampling plans lack appropriate adjustments when auditing settings change.

Taken together, these studies reveal a high relevance of arbitrary error containment, requiring the need for a better understanding of the in-the-field application of sampling. We therefore examine the actual effect of the newly implemented advice by ISA 530 and the specific issues addressed by the standard on auditors' error projection decisions.

7.3.3 Hypothesis Development

Recent studies assume that observed auditor behavior is in direct conflict with the requirements of auditing standards. The legitimization of error containment might lead to

adverse changes in auditing practice, while at the same time strengthening sensitization towards corresponding risks.

Elder et al. (2013) raise the question of whether auditors' decisions to isolate errors might be related to the formal acknowledgement of anomalies in auditing standards. As ISA 530 is the only standard to acknowledge error containment, our research focuses on two major properties of revealed errors contained in the standards' guidance: error frequency and the presence of containment information. Regarding the representative heuristic, we expect the guidance to strengthen the sensitivity of participants towards the statistical foundation of the necessity to project errors as well as the risks related to error containment. From a statistical point of view, neither characteristic should have any impact on projection decisions. However, with the standard initially introducing error containment, participants might as well pick up on the normative possibility. This leads us to our first hypothesis.

H1: Participants provided with ISA 530 guidance will project more errors than those making intuitive containment judgments.

ISA 530.12 states that, "The auditor shall investigate the nature and cause of any deviations or misstatements identified, and evaluate their possible effect on the purpose of the audit procedure and on other areas of the audit", creating a valid containment scenario if additional tests are perceived as sufficient and effective. For example, if a false entry of a credit note, caused by a certain employee, is revealed, this observation might lead to an examination of all credit notes issued by that accountant in the year under audit. This additional examination produces information regarding the nature and cause of the error, favorably justifying the validity of containing it to a sub-set of the original population under audit. However, other accountants might have made the same mistake. Hence, this strategy cannot be satisfying for a definite audit conclusion. Dusenbury et al. (1994) define containment information as context-specific information, which is usually gained by examining previously unaudited subpopulations of transactions which might be the only places the potentially isolatable error recurs. As previous studies suggest the latter exercise to be common practice, we expect the availability of containment information to increase the probability of not projecting an error and explicit guidance on containment information to mitigate the tendency to contain errors.

H2: The presence of containment information will be associated with lower error projection rates, with the effect being mitigated when ISA 530 guidance is present.

Referring to ISA 530.A17, we separately manipulate error frequency within subjects to address individual perceptions of similarity between causes of errors, as containment information might have only an indirect effect on containment decisions. When error causes are compared to known prototypes of possible causes, more frequent errors (errors commonly encountered) should lead to larger similarity perceptions than less frequent errors (e.g. fraudulent actions). As similarity perceptions are based on relative frequencies (Kahneman and Tversky 1973; Tversky 1977), we expect the probability of containment to be lower when an auditor perceives an error to be more common, hence being of higher frequency.

H3: Errors with a perceived higher frequency will be projected more often than errors perceived to occur less frequently, regardless of the presence of ISA 530 guidance.

7.4 Research Method

7.4.1 Participants

A total of 80 students in advanced semesters of a Master's programme in Finance and Accounting at a large public university in Germany participated in our experiment. All participants were enrolled at a senior-level auditing class. Because of their close proximity to graduation, these students were assumed to be reasonable surrogates for entry-level auditing associates. As interviews during pretests indicate that the entry-level auditing associates are usually responsible for the selection and evaluation of auditing samples, but do not receive formal training in bias avoidance techniques, we argue that these individuals are suitable for the task.[188] Conducting our tests with students allows avoiding biases due to familiarity with firm-specific auditing approaches, which may unconsciously predefine particular sampling plan strategies. All participants indicated at least a moderate knowledge of audit sampling theory and a basic knowledge of statistics. By the time the experiment was conducted, audit sampling theory was being discussed in their course. Incentives were set, as participants gained bonus points for their end-of-the-year exams; however, participation in the research project was voluntary and not part of the class fulfilment. Furthermore, students were able to pass the exam with the highest grade even without participating in the experiment.

[188] Prior research also indicates that biases attendant to representativeness found in students' judgments are just as present in the judgments of experienced professional auditors, who tend to overgeneralize in familiar auditing situations (Smith and Kida 1991).

7.4.2 Research Design and Experimental Procedure

To ensure students' familiarity with required theory, the experiment was conducted shortly before the year-end exam. Random assignment of the instrument resulted in a balanced distribution of observations across experimental conditions. The adoption of ISA 530 in Germany will be effective for audits of financial statements for periods beginning on or after December 15th, 2016. We conducted the experiment in 2017, when participants were likely not aware of the normative changes examined, as these will come into effect for the audit season in 2018. We focus on the impact of the guidance, especially across two factors that are essential in any error evaluation (Hermanson 1997; Dusenbury et al. 1994) while being explicitly considered in the wording of ISA 530: the amount of containment information available, and the error frequency, which should reflect the similarity of the cause of the observed error to other potential causes of errors.

The experimental task was a modified version of the instrument used by Hermanson (1997) and Dusenbury et al. (1994), which was originally developed by Burgstahler and Jiambalvo (1986). We designed eight cases comprising various errors. Participants evaluated each error by rating the necessity of error projection. To ensure accordance with indisputable error scenarios, some cases served as control scenarios. Participants were asked to rate each error independently. They were instructed to assume being the auditor in charge of the annual audit of a medium-sized entity. Background information on the company's performance and risk assessments was given, and participants' attention was drawn to the large balance of accounts receivable.[189] The narrative section also included several distracter cues. We defined the population to be represented by 1,000 individual debt accounts. A sample with a size of $n = 50$ was randomly selected and examined in substantive tests of details. As results from Elder and Allen (1998) and Peek et al. (1991) indicate, auditors are not likely to make the effort to contain an error if the projected amount is immaterial; hence, error projections in our cases lead to an estimate for total error exceeding tolerable error, resulting in non-trivial consequences. An essential element of error resolution is the consideration of additional courses of action, which can include (1) performing tests in specific areas (containment strategy), (2) increasing sample size to strengthen validity, (3) adjusting the balance, and (4) issuing a qualified audit

[189] We assume that nowadays accounts receivable are the most common audit field to provide a basis for our experiment, owing to reasonable homogeneity and uniform audit procedures performed in the field (Elder and Allen 1998).

opinion (Arens and Loebbecke 1999). To draw the participants' focus onto (1), we fixed all other actions within the case narrative.

A 2 x 2 x 2 mixed between- and within-subjects factorial design with fixed factors was used to test our hypotheses. We manipulated the variable of interest – presence of *GUIDANCE* – between subjects. Participants were randomly assigned to the two *INFORMATION* conditions, with further information on error containment procedures being either present or absent. *FREQUENCY* was a within-subjects variable, manipulated by including four cases with high-frequency errors and four cases with low-frequency errors. The instrument was pre-tested in a pilot study. Five academics and six auditors (three seniors, two managers, one partner) completed the study in an average of 52 minutes. To control for order effects, we conducted sensitivity analyses by incorporating various combinations of cases in our ANCOVA model, showing that there was no need for randomization of the case appearance. A debriefing of the pilot participants revealed that the task and manipulations were realistic and understandable. However, several suggestions were made to further improve the realism of the task and the individual cases. All suggestions were incorporated into the final instrument. The average time taken by the participants to complete the final instrument was 30 minutes.

7.4.3 Model

The following linear regression model was used to test the hypotheses:

PROJECTION = f(*GUIDANCE, INFORMATION, FREQUENCY, EXPERIENCE, KNOW_STAT*)

Table 7.1: Model Predictions and Description of Variable Coding

Variable	Predicted Sign	Coding Description
Dependent Variable:		
PROJECTION		= 1 to 7 rating (where 1 = contain error and 7 = project error)
Independent Variables:		
GUIDANCE	+	= 1 if ISA 530 provided = 0 otherwise
FREQUENCY	+	= 1 if common error (high frequency) = 0 otherwise
INFORMATION	-	= 1 if additional information provided = 0 otherwise
Control Variables:		
EXPERIENCE	+	= 1 to 7 rating (where 1 = none and 7 = extensive)
KNOW_STAT	+	= 1 to 3 rating (where 1 = none and 3 = extensive)

7.4.4 Dependent Variable

We want to address the effects of error containment guidance and error-specific properties on auditors' decisions to project errors. Our dependent variable was therefore *PROJECTION*, reflecting the participants' tendency to project or contain an error. It is assumed that individual decisions about the necessity to project errors are made using similarity perceptions, which reflects the participants' matching of errors to other errors commonly revealed in financial statement audits. *PROJECTION* in our model is indicated on a seven-point Likert scale, ranging from one (do not project error) to seven (definitely project error). The task was repeated over eight trials.

7.4.5 Independent Variables

GUIDANCE. Elder et al. (2013) state that auditing standards play a central role in auditors' decisions. However, the presence of normative guidance can have adverse effects, as auditors might tend to circumvent expected behavior because of perceived inconveniences (Kachelmeier and Messier 1990; Ashton 1983). With ISA 530 being the only standard to formally address error containment, we expect the availability of this guidance to affect auditors' judgments. Participants in the guided condition received full wording of the corresponding ISA 530 passages, while the rest of the narrative was held constant. With respect to the representative heuristic, we expect the presence of ISA 530 to increase the probability of judgments being in accordance with normative perceptions, resulting in lower error projection rates.

INFORMATION. Auditors are more likely to modify their error projection decisions when audit evidence suggests that an error is likely to occur in a well-defined sub-population (Christensen et al. 2015; Durney et al. 2014; Wheeler et al. 1997). In line with Dusenbury et al. (1994), we define containment information as "information about whether the error is known to be confined to a sub-set of examined items", resulting in "examining previously unsampled sub-populations of transactions in which the error might plausibly recur". ISA 530 is the only standard that acknowledges such properties of audit samples; hence, presence of the guidance might play a significant role in auditors' decisions to contain errors. To test for the effect, *INFORMATION* was manipulated at two levels by altering the case narratives, comprising either an uninformative or an informative condition. Participants without the treatment received information on the extent and cause of an error, but no additional information on the investigation of a population subset.

FREQUENCY. Hermanson (1997) and Dusenbury et al. (1994) argue that common errors are more likely to be projected onto a population than irregularities, as common errors augment an auditor's perception of similarity to other errors usually revealed in an audit. ISA 530 directly addresses error projection combined with fraud investigations. Thus, we manipulated error frequency within subjects at two levels. Four cases implied high-frequency errors (pricing error, transaction to the wrong debtor's account, misfooting of an invoice, recording of false payment conditions), and four cases implied low-frequency errors (sale under uncommon terms, deliberate invoice misfooting, fictitious sale, misdirection of a confirmation of balances by the auditor). By using the rounds model as suggested by Burgstahler et al. (2000), we increase the number of observations obtained from

each participant and examine their understanding of the effects of error frequency and cause of the error, while holding constant all other knowledge and experience effects that may influence the decisions.[190]

7.4.6 Control Variables

EXPERIENCE. When professional judgment is required, experience is an important factor in determining the outcome of a decision process (Taylor and Dunnette 1974; Libby 1985). In line with Messier (1983) and Nanni (1984), we expect task-related experience to influence error projection decisions for two reasons. First, general experience in revealing errors might lead to a more accurate consideration of uniqueness, and hence will favor higher projection rates. Second, even in the event of uncommon errors and the presence of containment information, experienced auditors will be more sensitive to the consequences of containing an error, especially in a non-guided condition (Abdolmohammadi and Wright 1987). Following Bedard and Biggs (1981), *EXPERIENCE* is measured by the participants' indication of working experience with error containment on a seven-point Likert scale, ranging from one (no experience at all) to seven (very experienced).

KNOW_STAT. Grounded knowledge of statistical coherences favorably influences subjective probability assessments (Libby 1985; Dubé-Rioux and Russo 1988). Christensen et al. (2015) attribute auditors' difficulties in projecting errors to a lack of understanding of the underlying statistical sampling concepts and related technical demands. We therefore asked participants to self-assess their knowledge in statistics within auditing contexts. Following Winkler (1967a, 1967b) and Stael von Holstein (1972), we anticipate better statistical understanding to favor responses more consistent with statistical principles. Hence, we expect participants familiar with statistical methods to project more errors. With wording slightly altered from that of Blocher and Bylinski (1985), we measure statistical knowledge on three levels: no skills, moderate skills, and extensive skills.

[190] An earlier version of the instrument had control risk in line for an additional manipulation. As control risk is usually already considered in setting sample sizes (Kachelmeier and Messier 2001, Christensen et al. 2015) and is not expected to have a significant effect on error projection decisions (e.g. a weaker control structure does not lead to a greater likelihood of error projection, Hermanson (1997)), we did not manipulate control procedures and set control risk at low to moderate levels by including corresponding information in the narrative.

7.4.7 Cases

The cases incorporated in our instrument are derived partly from Burgstahler and Jiambalvo (1986) and Dusenbury et al. (1994). To strengthen the validity of our instrument and to test for auditor consensus, some cases were added, while some other cases were omitted. The final cases are shown in Table 7.2, with text in parentheses being provided only in the containment information scenario. Cases one to four represent high-frequency errors, cases five to eight represent low-frequency errors. Table 7.2 also shows our predictions and normative perceptions of error projection decisions.

Table 7.2: Error Scenarios (Cases) in the Instrument and Expected Error Projection Decision

Case 1 (Error should be projected; Expectation: Error will be projected)
(1) Account No. # 211 is overstated by € 2,556.70 due to a pricing error. (2) The client´s accountant Mr. Meier apparently charged the price for an item located just below the correct item in the price book. (3) The error was not detected by the normal review of invoice accuracy but was revealed by the auditor. [(4) Mr. Meier handles and books similar accounting transactions on a daily basis. (5) An expansion of sample size and the examination of 30 more price calculations by Mr. Meier did not reveal further misstatements.]

Case 2 (Error should <u>not</u> be projected; Expectation: Error will <u>not</u> be projected)
(1) Account No. # 217 is overstated by € 4,215.87. (2) A sale was posted to this account when it should have been (but was not) posted to Account # 271. [(3) An examination of the sales invoice and shipping document indicated that the amount represents a valid sale. (4) During the review of the confirmations of balances, no other cases with the same cause of error were revealed.]

Case 3 (Error should be projected; Expectation: Error will <u>not</u> be projected)
(1) Account No. # 593 is overstated by € 3,687.37 due to an error on a credit memorandum. (2) The client apparently misfooted the invoices related to the credit by € 3,687.37. [(3) In total, only 85 credit memos were issued during the year. (4) To evaluate the extent of the error uncovered, the auditor examined all credit memos and checked the amounts by footing the totals on supporting documents. (5) No additional footing errors were noted.]

Case 4 (Error should be projected; Expectation: Error will <u>not</u> be projected)
(1) Account No. # 947 is overstated by € 8,865.00. (2) During December, the client agreed to sell € 118.200,00 in merchandise to this customer with payment due in six months and 15 % p. a. interest. (3) An accounting clerk recorded the sale, including unearned interest in the selling price; hence, revenues and accounts receivable are overstated by 8,865.00. [(4) Discussions with client personnel indicate that all other sales were transacted with the normal terms of PAYMENT DUE ON DELIVERY and did not include interest charges. (5) A special exception was made for customer # 947, to obtain his business. (6) Numerous tests of sales transactions were conducted in samples by the auditor throughout the year. (7) No sales arrangements including terms other than PAYMENT DUE ON DELIVERY were noted.]

Case 5 (Error should be projected; Expectation: Error will <u>not</u> be projected)
(1) Account No. # 685 is overstated by € 3,095.00. (2) A 15 year old fork-lift truck had been shipped on consignment to a dealer in used equipment but, due to a misunderstanding, recorded as a sale using the excepted selling price. [(3) Discussions with client personnel indicated that no additional consignment sales were made. (4) An examination of the equipment records indicated that no other items were retired during the past year.]

Case 6 (Error should be projected; Expectation: Error will <u>not</u> be projected)
(1) Account No. # 41 is understated by € 2,945.85. (2) However, the customer confirmed the proper balance. (3) It was determined that a clerical employee, who was related to a senior employee of the client´s customer, deliberately misfooted sales invoices for the account and forged the initials of the superior who reviews all invoices for accuracy. [(4) The auditor has determined that the employee was temporary and examined all invoices processed by the clerical employee for customer # 41 and revealed no additional errors. (5) Additionally, the auditor examined a sample of invoices for other customers processed by the temporary clerical employee and found no additional errors.]

Case 7 (Error should <u>not</u> be projected; Expectation: Error will <u>not</u> be projected)
(1) Account No. # 227 is overstated by € 11,456.72 due to a fictitious sale submitted by a salesman. (2) It was revealed that this was part of the salesman's effort to boost his sales near the end of the first quarter, in order to achieve personal objectives. [(3) The auditors' investigation revealed that the employee, who began working for the client in mid-January, was dismissed at the beginning of April due to customer complaints, but the client had not uncovered his fraud by then. (4) The auditor examined any sales made by this salesman during the year and identified no further fictitious sales. (5) All amounts were verified by documentation and by confirmations of customers.]

Case 8 (Error should be projected; Expectation: Error will be projected)
(1) Account No. # 329 is overstated by € 2,625.30. (2) By referring back to the sampling plan, the auditor discovered that due to his own error (in relating random numbers to sample account numbers), account No. 329 instead of account No. # 929 had mistakenly been requested. (3) Hence, account No. # 329 was not part of the initial sample. (4) The auditor promptly mailed a confirmation to account No. # 929, and the confirmation was returned without exception. [(5) By a repeated comparison of the account balance list with the requested confirmations of balances, the auditor ensured that the error occurred only once, and that no further confirmations were misdirected.]

7.5 Results

7.5.1 Manipulation Checks

Prior to data analyses, we ensured containment information was not attributed to the cause of an error, by confirming that participants correctly identified error causes. Overall, 80.1% of participants' recognition of error causes was correct. Additionally, we ensured that participants understood the cases and appreciated manipulations by having them complete manipulation checks within the debriefing questionnaire. Of all the participants, 100% correctly indicated the presence of containment guidance in the guidance-present condition, whereas 60% indicated that no further guidance was present in the intuitive condition, indicating a successful manipulation of *GUIDANCE*. The perceived amount of containment information was rated on a 7-point scale with 1 denoting none and 7 denoting very extensive. The mean (standard deviation) rating for *INFORMATION* in the *INFORMATION*-absent environment was 3.9 (1.2), whereas in the *INFORMATION*-present environment it was 4.7 (1.4). The difference is statistically significant ($t = 3.73$; $p < 0.001$),

indicating a successful manipulation of *INFORMATION*. For each trial, participants were asked to assess the perceived *SIMILARITY* to errors usually revealed in an audit, rated on a 7-point scale ranging from 1 (not similar at all) to 7 (very similar). The mean (standard deviation) rating for *SIMILARITY* in the low-frequency condition was 3.8 (1.16), and 4.8 (0.90) in the high-frequency condition. The difference is statistically significant at (t = 6.02; $p < 0.001$), indicating a successful manipulation of *FREQUENCY*.

7.5.2 Descriptive Statistics

Descriptive statistics for the eight treatment conditions are shown in Table 7.3. Panel A shows means and standard deviations for the dependent variable *PROJECTION*. Panel B of Table 7.3 shows means and standard deviations by experimental condition. First and foremost, *GUIDANCE* obviously does not uniformly affect *PROJECTION* means in the presence of *INFORMATION*. However, mean error projection rates are lower when *GUIDANCE* is provided, and *INFORMATION* is absent, raising the question of whether the intended effective direction of ISA 530 can be achieved. Most noticeably, when *INFORMATION* is present, *PROJECTION* means are considerably lower than in information-absent conditions. When *INFORMATION* is present, *PROJECTION* means are lower in low-*FREQUENCY* conditions. However, in the *INFORMATION*-absent conditions, *FREQUENCY* does not affect *PROJECTION* means.

Table 7.3: Descriptive Statistics for *PROJECTION*[a]

Panel A: Descriptive Statistics by Cell (cell n = 160)

		GUIDANCE[d]	
INFORMATION[b]	***FREQUENCY***[c]	**Present Mean (Std. Dev.)**	**Absent Mean (Std. Dev.)**
Present	High	3.4 (1.6)	3.6 (1.9)
	Low	3.0 (1.3)	2.4 (1.1)
Absent	High	4.1 (1.8)	4.7 (1.6)
	Low	4.2 (1.0)	4.6 (1.1)

Panel B: Descriptive Statistics by Condition

GUIDANCE		*INFORMATION*		*FREQUENCY*	
Present Mean (Std. Dev.)	**Absent Mean (Std. Dev.)**	**Present Mean (Std. Dev.)**	**Absent Mean (Std. Dev.)**	**High Mean (Std. Dev.)**	**Low Mean (Std. Dev.)**
3.6 (1.5)	3.9 (1.7)	3.1 (1.6)	4.4 (1.4)	4.0 (1.8)	3.6 (1.4)
n = 78	n = 82	n = 78	n = 82	n = 80	n = 80

[a] *PROJECTION* is measured on a seven-point Likert scale.
[b] *INFORMATION* is manipulated between subjects at two levels: Present and absent.
[c] *FREQUENCY* is manipulated within subjects at two levels: high-frequency and low-frequency.
[d] *GUIDANCE* is manipulated between subjects at two levels: Present and absent.

7.5.3 Main Results

The homogeneity of variances was assessed using Levenes' Test, (p = .4209). We tested our hypotheses using a 2 x 2 x 2 ANCOVA design with two between-subjects factors, *GUIDANCE* and *INFORMATION*, as well as one within-subjects factor, *FREQUENCY*. Panel A of Table 7.4 shows the full model results including our control variables *EXPERIENCE* and *KNOW_STAT*.

Hypothesis H1 states that participants provided with ISA 530 guidance will project more sample errors than participants not provided with the guidance. The main effect for *GUIDANCE* on *PROJECTION* is found to be not significant (see Table 7.4, Panel A). The main result we observe here is that the judgment of whether to contain or project an error cannot be driven by additional guidance alone, hence ISA 530 does not affect participants' perception of similarity. Thus, we must reject our H1.

Panel A of Table 7.4 also shows the results of our H2 and H3. Hypothesis H2 states that the presence of containment information will influence auditors' decisions on error containment in such a way that error projection rates will be higher when containment information is absent as compared to environments where containment information is present, with the effect being stronger in the presence of *GUIDANCE*. Likewise, hypothesis H3 indicates that error projection rates will be higher when errors are perceived to be more frequent, with the effect being stronger in the presence of *GUIDANCE*. The main effects of both variables, *INFORMATION* and *FREQUENCY,* on *PROJECTION* are significant with p-values of 0.000 and 0.082, respectively. Thus, hypotheses H2 and H3 cannot be rejected, indicating that both variables override a potential main effect of *GUIDANCE*.

Table 7.4: Test of H1, H2 and H3

Panel A: Full Model[a]

Source	SS	df	F	p-value
GUIDANCE	0.74	1	0.34	0.558
INFORMATION	62.79	1	29.22	0.000
GUIDANCE x INFORMATION	6.16	1	2.87	0.093
FREQUENCY	6.60	1	3.07	0.082
GUIDANCE x FREQUENCY	1.99	1	0.93	0.337
INFORMATION x FREQUENCY	7.41	1	3.45	0.065
GUIDANCE x INFORMATION x FREQUENCY	0.97	1	0.45	0.503
EXPERIENCE	0.23	1	0.11	0.745
KNOW_STAT	6.19	1	2.88	0.092

Panel B: Simple Effects of *INFORMATION* at Different Levels of *GUIDANCE* and *FREQUENCY*

		Effect of *INFORMATION*[b]	
GUIDANCE	***FREQUENCY***	**F**	**p-value**
Present	High	1.55	0.215
	Low	5.81	0.018
Absent	High	5.45	0.021
	Low	23.91	0.000

Panel C: Simple Effects of *GUIDANCE* at Different Levels of *INFORMATION*

	Means		Effect of *INFORMATION*	
GUIDANCE	**Present**	**Absent**	**F**	**p-value**
Present	3.20	4.18	8.38	0.004
Absent	2.98	4.63	24.84	0.000

Panel D: Effects of *INFORMATION* at Different Levels of *FREQUENCY*

	Means		Effect of *INFORMATION*	
FREQUENCY	**Present**	**Absent**	**F**	**p-value**
Low	2.69	4.44	28.23	0.000
High	3.50	4.42	7.80	0.006

[a] Between-subjects tests are performed for *INFORMATION* and *GUIDANCE*; within-subjects tests are performed for *FREQUENCY*.
[b] Between-subjects tests are performed for *INFORMATION*.
PROJECTION is measured on a seven-point Likert scale.
GUIDANCE is manipulated between subjects at two levels: Presence or absence.
INFORMATION is manipulated between subjects at two levels: Presence and absence.
FREQUENCY is manipulated within subjects at two levels: high-frequency and low-frequency.

Although we cannot observe a main effect for GUIDANCE, we find a significant interaction effect of *GUIDANCE* with *INFORMATION* (p = 0.093) in our mixed model (see Table 7.4, Panel A). In addition to the observed main effects, we also find a significant interaction effect of *INFORMATION* with *FREQUENCY* (p = 0.065). We investigate this further by performing a simple effects analysis of *GUIDANCE* at different levels of the two other independent variables, *RISK* and *PROCEDURES*. This analysis is shown in

Panel B of Table 7.4. We find that *INFORMATION* is significant at all levels of the other independent variables, except when *GUIDANCE* is present and *FREQENCY* is high (*p*-values of 0.215, 0.018, 0.021 and 0.000). However, Panel A of Table 7.3 shows that the effects are not all in the hypothesized direction. First, the presence of *GUIDANCE* results in lower *PROJECTION* rates when *INFORMATION* is absent. Thus, we find strong support for our hypothesis that the presence of ISA 530 affects the tendency of individuals to contain errors in accordance with the representative heuristic: the normatively provided opportunity to contain errors drives participants to assume this option valid, even when there is no effective evidence for its applicability. Second, *FREQUENCY* does affect participants' error projection decisions, especially when additional containment information is provided. Thus, only when *INFORMATION* is present participants decide to project more common errors as opposed to uncommon errors. We examine the observed significant effects in more detail within our additional analyses.

7.5.4 Controls

Panel A of Table 7.4 shows that only the *KNOW_STAT* main effect is significant, with a *p*-value of 0.092. As expected, more in-depth knowledge of statistics leads to higher error projection rates. We conclude that a move towards normatively accepted projection rates can only be achieved if an auditor possesses a working knowledge of the statistical concepts underlying audit sampling theory, e.g. by taking part in statistics courses during university or by taking the auditors' exam.

By contrast, the main effect for *EXPERIENCE* as a proxy for working knowledge with error containment is not significant in our model. Thus, any participant, regardless of how long they have worked in the auditing environment and their experience of error containment, might evaluate and resolve sampling errors in a normatively non-defensible way and run the risk of drawing invalid conclusions about an underlying population.

Table 7.5: Participants' error *PROJECTION* ratings at different levels of *INFORMATION*

Cases (n = 160)	Mean *PROJECTION* Rating: *INFORMATION* Absent (n = 82)	*INFORMATION* Present (n = 78)	p-value	*EXPERIENCE*	*KNOW_STAT*
High-frequency Errors					
1. pricing error	4.1	2.5	0.000	0.073	n.s.
2. posting error	3.9	3.1	0.002	n.s.	0.053
3. misfooted credit memo	4.8	2.5	0.000	n.s.	n.s.
4. unearned interest	4.5	3.1	0.000	n.s.	n.s.
Low-frequency Errors					
5. consignment sale	4.2	3.2	0.001	n.s.	n.s.
6. fraudulent misfooting	4.5	3.8	0.023	n.s.	n.s.
7. fictitious sales	4.5	3.5	0.002	n.s.	n.s.
8. misdirected confirmation	3.9	3.7	0.197	0.086	n.s.

To examine these indicators in more detail, we exclude *FREQUENCY* from our full model and perform ANCOVAs for each of our cases separately, examining the individual effects of information and the control variables. The basic results can be found in Table 7.5, showing that either *EXPERIENCE* or *KNOW_STAT* is significant only for those cases where our normative assumption of the projection decision meets our expectations towards participants' behavior (see Table 7.2). Hence, we conclude that misconceptions and audit failures are present more often in environments requiring a more complex error resolution, and that experience and statistical knowledge do not necessarily moderate this conduct.

However, we must interpret these results cautiously, as the basis for our measurements of *EXPERIENCE* and *KNOW_STAT* is provided only by the participants' self-assessment, and we therefore cannot guarantee the avoidance of a distortion in the recorded self-evaluation. As a result, neither of these variables may reflect the concrete understanding of the task fulfilment or task-specific expertise.

7.5.5 Additional Results

First, as previously noted, we find a significant interaction of *INFORMATION* with *GUIDANCE* with a *p*-value of 0.093 (see Panel A of Table 7.4). To investigate this further, we perform a simple effects analysis of *GUIDANCE* at different levels of *INFORMATION*. This analysis is shown in Table 7.4, Panel C, and illustrated in Figure 7.1. We can see that when *INFORMATION* is present, *GUIDANCE* leads to significantly higher projection rates. Interestingly, the effect runs in another direction, when *INFORMATION* is absent, implying that participants are less sensitive to flawed audit procedures in the

presence of ISA 530 in non-informative environments. Hence, the guidance comprised in ISA 530 prevents auditors from projecting errors, especially when no containment information is present. In other words, ISA 530 strengthens audit quality in the most typical audit scenarios, that is, when *INFORMATION* is present, owing to the performance of additional auditing procedures, but diminishes audit quality in riskier audit environments, that is, when *INFORMATION* is absent.

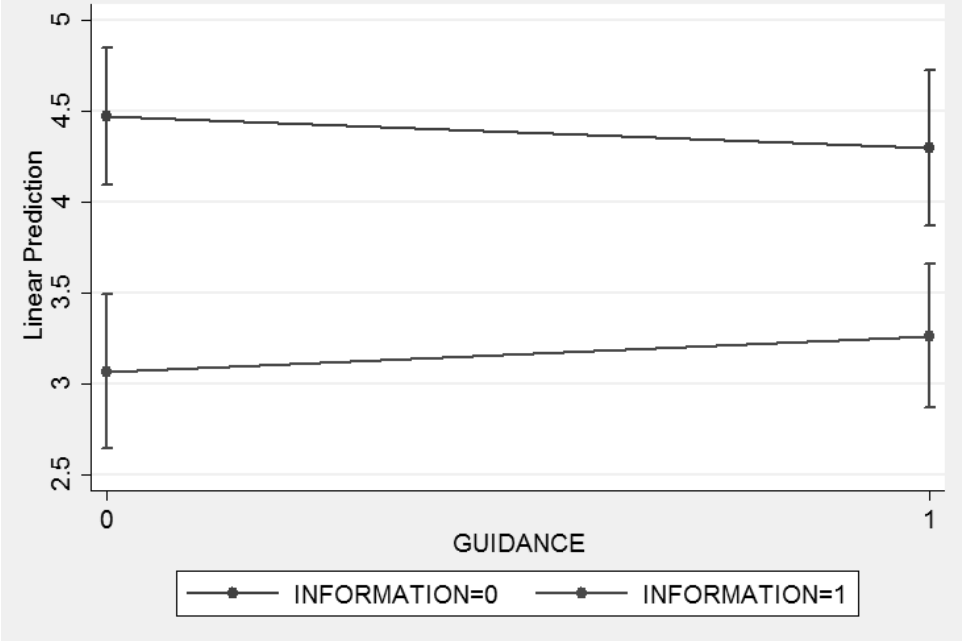


Figure 7.1: Adjusted predictions of *GUIDANCE* and *INFORMATION* with 95% CIs

Second, Panel A of Table 7.4 also shows that *INFORMATION* interacts significantly with *FREQUENCY* ($p = 0.065$). We therefore perform a repeated measures ANOVA, examining the effect of the within-subjects variable *FREQUENCY* at each level of *INFORMATION* (see Panel D of Table 7.4). As our illustration in Figure 7.2 shows, *FREQUENCY* comes strongly into effect when *INFORMATION* is present. This outcome is quite surprising, as each error and its cause are accurately predefined in our cases for all participants, in both the presence and the absence of additional containment information. In line with various psychological studies (Hansen and Waenke 2010; Freitas et al. 2004; Schooler et al. 1986), we argue that individuals become more confident of their own decisions when more concrete information is given, regardless of the substantial relevance of this information. As more information leads to more complexity and cognitive strain, the application of the representative heuristic might become more likely and error containment be perceived to be justifiable.

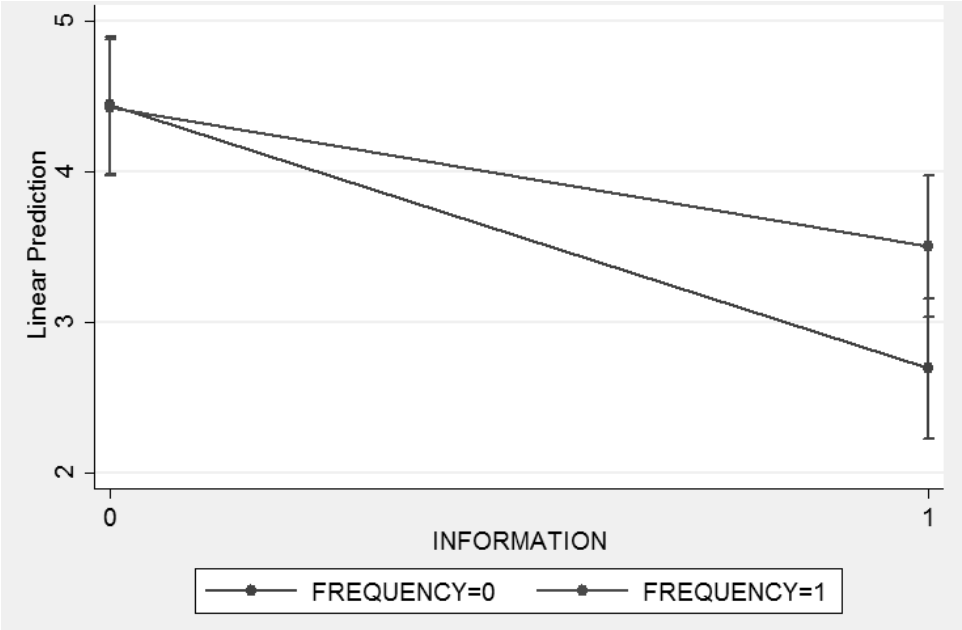


Figure 7.2: Adjusted predictions of *INFORMATION* and *FREQUENCY* with 95% CIs

Third, we intend to provide a link between error containment judgments and auditors' perceptions of similarity, as a central phenomenon in decision heuristics is that a perceived uniqueness of errors can prevent similarity matches (Tversky 1977; Burgstahler and Jiambalvo 1986). To test for a link between error containment and error cause, participants were required to evaluate each error in respect of its uniqueness, by assessing its similarity to other errors usually revealed. Similarity was assessed on a seven-point scale ranging from one (not similar) to seven (very similar). Participants were also asked to consider qualitative aspects, as the nature and cause of an error can override quantitative findings, e.g. a dedication to related party transactions or fraud. Thus, each projection decision was followed by an identification of the cause of the error, and a similarity assessment of the chosen cause to other causes of errors commonly revealed in the field.

Table 7.6 gives the mean values of *SIMILARITY* for both error types (high- and low-frequency errors), as well as the results of a paired samples t-test.[191]

Table 7.6: Paired samples t-test

Participants' Perceived *SIMILARITY* in Low- and High *FREQUENCY* Conditions

FREQUENCY						
Mean for Low	**Mean for High**	**Mean Difference**	**Std. Deviation**	**df**	**t**	**Sig.**
3.8	4.8	1.0	1.15	158	6.02	0.01

n = 160
SIMILARITY is measured on a seven-point Likert scale.
FREQUENCY is manipulated at two levels: Low and High.

Regardless of the presence of *GUIDANCE* and the availability of *INFORMATION*, *SIMILARITY* ratings within the two *FREQUENCY* conditions differ strongly, whereas *PROJECTION* decisions do not significantly differ for either type of error. *SIMILARITY* ratings are consistently higher for high-frequency errors than they are for low-frequency errors (4.9 and 4.0, respectively), whereas *PROJECTION* decisions do not reflect these perceptions. Detailed analyses show a mean for *PROJECTION* of 3.6 for high-frequency errors and 4.0 for low-frequency errors (see Table 7.7). We cannot explain this observation conclusively. However, the effect is driven mostly by the *INFORMATION*-present conditions, meaning that participants tended to contain more high-frequency errors, even when they were obviously conscious of the similarity to other errors. By containing these errors, participants chose to selectively ignore the assumption that sample observations were helpful in gathering information about the remainder of the population.

Table 7.7: Comparison of participants' error *PROJECTION* and *SIMILARITY* ratings

	Mean *PROJECTION* Rating	**Mean *SIMILARITY* Rating**
Case 1	3.3	5.0
Case 2	3.5	5.1
Case 3	3.7	4.8
Case 4	3.8	4.7
Case 5	3.7	4.4
Case 6	4.2	3.2
Case 7	4.1	4.0
Case 8	3.8	4.3

[191] Because our observations in the two samples are not independent of each other, paired (dependent) t-tests are used in this analysis.

Finally, we examine auditor consensus for all cases. Cases 2 and 8 were added as controls to assess consistency between participants' error projection decisions. However, agreement was not as high as we expected. Table 7.5 shows that in Case 8 only, *INFORMATION* had no significant effect and that the mean shift between conditions was, as expected, very low. For Case 2, the mean *PROJECTION* rating is the same as for Case 8, although the error is much more likely to arise from typical accounting errors. Although the mean shift is slighter than for the other high-frequency errors, *INFORMATION* has a significant effect on the participants' error projection decision for Case 2, when in fact it should have no influence. Hence, we conclude that a substantial percentage of auditors treated undoubted errors inappropriately, probably because of a lack of understanding of the consequences of their decisions, as error causes were identified correctly in the majority of all trials. These observations leave room for future studies to investigate personal and firm-specific effects of error containment strategies and the isolation of representative errors, which we could not control because of the background of our participants.

7.5.6 Discussion

The main result we observe is that an accurate evaluation and resolution of sampling errors is highly dependent on the individual perception of the validity of error containment. Regardless of guidance on error containment, participants tended to contain errors that were actually representative of an underlying accounting population, resulting in a lot of room for decision biases. Our observations show that these biases cannot be easily mitigated by basic normative guidance. While providing the option of treating errors as anomalies, ISA 530 fails to improve participants' performance, but rather induces them to initially contain errors. Our findings hold for two levels of containment information (presence vs. absence) and for two levels of error frequency (high vs. low). This supports our argument that biases in audit sampling can only be mitigated when an auditor is reminded in more depth of general statistical principles and the implications of their violation, which requires a more precise normative appreciation of the problem. In line with Dusenbury et al. (1994), we find that in the presence of containment information errors were significantly less often projected than in the absence of such information. As this scenario represents typical auditing situations, our results renew concerns about the estimated number of unknown cases in the auditing profession in which error containment is performed arbitrarily (Fay et al. 2015; Christensen et al. 2015) and highlight the need for a prevention of such action. Error frequency is also a significant factor. If errors are

perceived to be uncommon, the probability of error containment tends to be higher. However, uncommon errors (e.g. irregularities due to incidents of fraud) are more troublesome, as auditors seem to be easily convinced that these errors justify containment scenarios when in fact they indicate unusual and sometimes fraudulent circumstances.

Our results strongly support the representative heuristic theory. Despite the fact that error containment is only justifiable in extremely rare cases, our participants were overly optimistic that quite a few errors are unique. It is important to consider that additional normative guidance is not meant to justify increases in containment decisions. However, there are fortunately attenuating circumstances, as it will be seldom the case that an auditor in the field fails to project an error and at the same time the account balance is materially misstated. Yet, this indication is more problematic than satisfying, as auditors cannot calculate the chance of such an occurrence if the initially provided audit evidence is too vague to draw valid conclusions about entire populations.

7.6 Conclusions, Limitations and Future Research

Our study provides insights into error containment, which today is one of the major concerns in audit sampling (Fay et al. 2015; Christensen et al. 2015; Elder et al. 2013). ISA 530 requires that an auditor should estimate total population error by projecting individual errors to the account balance. Failure to project errors degrades this estimate, causing auditors to accept materially misstated accounting populations. We find strong evidence that sampling plans regularly do not follow the structures they are intended to. Additionally, we observe that a revision of auditing standards specifically addressing known concerns does not make auditors follow them unconditionally. Regardless of the presence of containment guidance and the frequency of errors, participants in our experiment failed to quantify error estimates by projecting them to the population, even if the errors were representative of and proxy for other accounting errors. Quite the opposite, participants tended to treat seemingly unique errors as anomalies, resulting in exceptions from error projection necessities.

Our results imply that concerns in the evaluation and resolution of sample errors persist. In line with Burgstahler et al. (2000) we find that a non-trivial portion of errors is, apparently, not projected. Hence, ISA 530 guidance on error containment seems too tentative and does not commit the addressees to appreciating statistical requirements rather than relying on intuitive decision making. As auditors today tend to under-audit rather than

over-audit (Christensen et al. 2015), this amplifies the risk associated with containing errors. The adverse consequences might lead to an increasing probability of accepting materially misstated accounting populations that would be rejected if an adequate estimator were used. Monitoring controls and additional training are the only possible solutions to these concerns.

Our study contributes to the growing body of behavioral research in auditing by extending enquiries into auditors' decision making in the resolution of revealed errors – looking at their tendency to accept increased detection risk by performing error containment. Moreover, we clarify the effect of changing audit standards from efficacy toward a greater judgmental consensus. Hence, our results offer important practical implications for auditing firms and regulators. Auditing firms should be aware of the observed naturalness of error containment practice as well as the deduced risks and should contemplate counteracting these flaws by implementing meaningful internal guidance and monitoring systems. Moreover, regulators can benefit from the findings we present here, as they might find the current wording of audit standards in need of improvement.

Our study is subject to some limitations associated with experimental design and procedure. Beside a possible lack of task realism, participants were also not selected at random. Consultation in an auditing team was not possible, hence, error resolution strategy could not be discussed. These conditions may have led our study to overestimate the extent to which auditors fail to project sample errors. The results may also be limited owing to a possible imprecision in the measurements of control variables, as task-related experience and statistical knowledge were assessed by the participants themselves, resulting from time and anonymity constraints. Our interpretations of the results discussed should be considered with these limitations in mind.

To summarize, we believe that the investigation of heuristics in terms of failing to project errors and the implications for audit risk is an important enrichment of the existing behavioral literature. We encourage future research on auditors' decisions when resolving misstatements and their tendency to bypass normative guidelines, by testing the effect of alternative strategies directed toward the mitigation of containment scenarios.

7.7 Appendix: Instrument

Note: Each participant received the following eight cases. Cases one to four include common ("high-frequency") errors; cases five to eight include relatively uncommon ("low-frequency") errors. All questions following the first case were repeated for all other cases. Italicized sentences were present in the *GUIDANCE*-present condition only. Sentences in square brackets were included in the *INFORMATION*-present condition only.

Welcome!

Thank you for participating in our research study on the efficacy of audit regulation in the environment of audits of financial statements. For the purpose of our research, we will survey individuals performing audits of financial statements and those who already focus their tenure throughout their study courses.

As participating students get bonus points for the year-end exam, you have the possibility to note your matriculation number at the end of the survey.

You will perform the case study from an auditor's perspective. Succeeding the introduction, you will receive general information about the annual financial statements for the fiscal year 2016 of the fictional company "Cookie Products Ltd.". Subsequently, you will obtain a narrative on eight specific issues of audit procedures, which require your assessment. We ask you to take the perspective of the auditor of „Cookie Products Ltd." throughout the case study. The time needed to read all necessary information and to finalize the case study will be about 30 minutes.

Please note that only before processing the first case you have the possibility to go back and read the introductory text on the following slide again. This possibility will be blocked as soon as you reach the second case. Furthermore, it is not possible to restart the survey.

Please reply honestly to any questions in the survey. There are no "right" or "wrong" answers. We are simply interested in observing your true behavior. All answers will be kept confidential.

We gratefully acknowledge your support!

page turn

COOKIE PRODUCTS LTD.

Assume you are auditor in charge for the audit of the annual financial statements of Cookie Products Ltd., a manufacturer of kitchen appliances. The company generated constant positive outcome throughout the past years. From the audit of internal control over financial statements as well as from analytical audit procedures there is no indication for an increased audit risk.

Your audit attention is directed towards the T€ 1.500 balance of accounts receivable. This amount represents 1,000 individual balances from debt accounts with numbers # 1 to # 1000.

An audit sample of individual debtor balances with a sample size of $n = 50$ was selected from all accounts receivable. 50 randomly selected customers were therefore asked to confirm year-end balances by written confirmations.

Within the audit, various misstatements were revealed by the confirmations. Necessarily, those misstatements need to be appraised in the following.

Your task for the following **8 individual cases** is to assess the **need of error projection of individual misstatements to the underlying population** of all accounts receivable. Please rate each error on its own and separately from all other errors, so as each error was revealed in a different audit.

The following **premises pertain for all individual cases**:

- In case you decide to extrapolate a misstatement to the underlying population (**error projection**): The extrapolation is calculated proportional to the audit scope. It is assumed that in this case, the materiality limit of T€ 40 for accounts receivable is always exceeded. That implies the necessity to extend the audit scope as well as a subsequent correction of the misstatement by a booking in the clients' records, and possibly the issuance of a qualified audit opinion.
- In case you decide not to extrapolate a misstatement to the underlying population (**error containment**): The individual misstatement will be corrected by a booking in the clients' records and has no further impact on audit procedures or the issuance of an unqualified audit opinion.

page turn

*Please note: In accordance with ISA 530, it is generally possible to classify an error detected within an audit procedure using audit sampling as a so-called **anomaly**. Hence the error **will be excluded from the error projection to the population**. While dealing with the cases, please take into account the following excerpts from ISA 530:*

*An **anomaly** is a misstatement or deviation that is **demonstrably not representative** of misstatements or deviations in a population.*

*In the **extremely rare circumstances** when the auditor considers a misstatement or deviation discovered in a sample to be an anomaly, the auditor shall obtain a **high degree of certainty that such misstatement or deviation is not representative of the population**. The auditor shall obtain this degree of certainty by **performing additional audit procedures** to obtain sufficient appropriate audit evidence that the misstatement or deviation does not affect the remainder of the population.*

*In analyzing the deviations and misstatements identified, the auditor may observe that many have a **common feature**, for example, type of transaction, location, product line or period of time. In such circumstances, the auditor may decide to **identify all items in the population** that possess the common feature and **extend audit procedures to those items**. In addition, such deviations or misstatements may be intentional, and may indicate the **possibility of fraud**.*

While processing the first case, it is possible to go back and view this information again.

page turn

Objective

Case 1

(1) Account No. # 211 is overstated by € 2,556.70 due to a pricing error. (2) The client´s accountant Mr. Meier apparently charged the price for an item located just below the correct item in the price book. (3) The error was not detected by the normal review of invoice accuracy but was revealed by the auditor. [(4) Mr. Meier handles and books similar accounting transactions on a daily basis. (5) An expansion of sample size and the examination of 30 more price calculations by Mr. Meier did not reveal further misstatements.]

1. For the previous case please indicate your tendency to either include the error in the sample estimate (error projection), or to exclude the error from the sample estimate (error containment). Consider the consequences of your decision explained in the introduction.

1 (definitely contain error = no error projection)	2	3	4	5	6	7 (definitely project error = no error containment)
O	O	O	O	O	O	O

2. How confident are you in this error projection / error containment decision?

1 (not confident at all)	2	3	4	5	6	7 (very confident)
O	O	O	O	O	O	O

3. Please give the number [number in parentheses, e. g. *(2)*] of the sentence which best describes the cause of the error:

   Sentence number _____ describes the cause of the error best.

4. Considering only the sentence that you selected as the cause of the error and no other information in the case, how similar is this cause to causes of other potential errors typically revealed during an audit?

1 (not similar at all)	2	3	4	5	6	7 (very similar)
O	O	O	O	O	O	O

_____*page turn*_____

Case 2

(1) Account No. # 217 is overstated by € 4,215.87. (2) A sale was posted to this account when it should have been (but was not) posted to Account # 271. [(3) An examination of the sales invoice and shipping document indicated that the amount represents a valid sale. (4) During the review of the confirmations of balances, no other cases with the same cause of error were revealed.]

_____*page turn*_____

Case 3

(1) Account No. # 593 is overstated by € 3,687.37 due to an error on a credit memorandum. (2) The client apparently misfooted the invoices related to the credit by € 3,687.37. [(3) In total, only

85 credit memos were issued during the year. (4) To evaluate the extent of the error uncovered, the auditor examined all credit memos and checked the amounts by footing the totals on supporting documents. (5) No additional footing errors were noted.]

page turn

Case 4

(1) Account No. # 947 is overstated by € 8,865.00. (2) During December, the client agreed to sell € 118.200,00 in merchandise to this customer with payment due in six months and 15 % p. a. interest. (3) An accounting clerk recorded the sale, including unearned interest in the selling price; hence, revenues and accounts receivable are overstated by 8,865.00. [(4) Discussions with client personnel indicate that all other sales were transacted with the normal terms of PAYMENT DUE ON DELIVERY and did not include interest charges. (5) A special exception was made for customer # 947, to obtain his business. (6) Numerous tests of sales transactions were conducted in samples by the auditor throughout the year. (7) No sales arrangements including terms other than PAYMENT DUE ON DELIVERY were noted.]

page turn

Case 5

(1) Account No. # 685 is overstated by € 3,095.00. (2) A 15 year old fork-lift truck had been shipped on consignment to a dealer in used equipment but, due to a misunderstanding, recorded as a sale using the excepted selling price. [(3) Discussions with client personnel indicated that no additional consignment sales were made. (4) An examination of the equipment records indicated that no other items were retired during the past year.]

page turn

Case 6

(1) Account No. # 41 is understated by € 2,945.85. (2) However, the customer confirmed the proper balance. (3) It was determined that a clerical employee, who was related to a senior employee of the client´s customer, deliberately misfooted sales invoices for the account and forged the initials of the superior who reviews all invoices for accuracy. [(4) The auditor has determined that the employee was temporary and examined all invoices processed by the clerical employee for customer # 41, and revealed no additional errors. (5) Additionally, the auditor examined a sample of invoices for other customers processed by the temporary clerical employee and found no additional errors.]

page turn

Case 7

(1) Account No. # 227 is overstated by € 11,456.72 due to a fictitious sale submitted by a salesman. (2) It was revealed that this was part of the salesman's effort to boost his sales near the end of the first quarter, in order to achieve personal objectives. [(3) The auditors' investigation revealed that the employee, who began working for the client in mid-January, was dismissed at the beginning of April due to customer complaints, but the client had not uncovered his fraud by then. (4) The auditor examined any sales made by this salesman during the year and identified no further fictitious sales. (5) All amounts were verified by documentation and by confirmations of customers.]

page turn

Case 8

(1) Account No. # 329 is overstated by € 2,625.30. (2) By referring back to the sampling plan, the auditor discovered that due to his own error (in relating random numbers to sample account numbers), account No. 329 instead of account No. # 929 had mistakenly been requested. (3) Hence, account No. # 329 was not part of the initial sample. (4) The auditor promptly mailed a confirmation to account No. # 929, and the confirmation was returned without exception. [(5) By a repeated comparison of the account balance list with the requested confirmations of balances, the auditor ensured that the error occurred only once, and that no further confirmations were misdirected.]

page turn

Further Questions

Please answer the following questions faithfully and honestly:

1. As far as you remember: Were individual accounts receivable balances of Cookie Products Ltd. tested in a sample?

1 (I don´t agree)	2	3	4	5	6	7 (I fully agree)
O	O	O	O	O	O	O

2. Which normative advices on error projection and error containment did you receive additionally to the general information about Cookie Products Ltd.?

   (O) I did not get any more advice for the case study
   (O) I got additional advice from ISA 530 on dealing with anomalies

3. [ONLY IF 2 = YES] How much did the additional advice from ISA 530 help you to get to a decision regarding error projection?

   (O) Helped me a lot
   (O) Helped me a little
   (O) Did not help me

4. How extensive was the provided information to get to a decision regarding error projection or error containment in the single cases?

1 (not present)	2	3	4	5	6	7 (very extensive)
O	O	O	O	O	O	O

5. How precisely did you read the introductory case study of Cookie Products Ltd. (Please reply honestly. None of the following answers is “right” or “wrong“; we are simply interested in your true behavior.)?

   (O) not read / bypassed
   (O) briefly / read across
   (O) fairly intensive (word for word / number for number)
   (O) very intensive (e. g. repeatedly, tried to memorize content)

6. How precisely did you read the individual cases and the information regarding causes of errors (Please reply honestly. None of the following answers is “right” or “wrong“; we are simply interested in your true behavior.)?

(O) not read / bypassed
(O) briefly / read across
(O) fairly intensive (word for word / number for number)
(O) very intensive (e. g. repeatedly, tried to memorize content)

7. Do you already have practical experience with audit sampling procedures?

(O) Yes
(O) No

8. Do you already have practical experience with error containment?

1 (No experience at all)	2	3	4	5	6	7 (Very experienced)
O	O	O	O	O	O	O

9. For how many years did you already work in audit environments?

_____ years

10. You are

(O) male
(O) female

11. How old are you?

_____ years

12. Which university degree do you already have?

(O) Bachelor
(O) Master
(O) Diploma
(O) other: _____

13. Which university degree do you strive for?

(O) Bachelor
(O) Master
(O) Diploma
(O) other: _____

14. How do you rate your personal skills regarding statistical methods in general?

(O) no skills
(O) moderate skills
[= general knowledge available, application of audit procedures mandatory]
(O) extensive skills
[= more in-depth understanding of auditing and application of audit procedures]

8 Conclusion

This dissertation deals with audit sampling theory, the revision of corresponding professional audit sampling standards, and the behavioral impact of the latter on auditors' decisions over sample size and error containment, as well as the cognitive challenges that auditors are faced with when applying non-statistical sampling procedures. The first paper (Chapter 2) reviews the development of audit sampling techniques and technological demands throughout the profession over the last four decades. It identifies the audit sampling fundamentals that drive the most disruptive problems in the field and evaluates changes planned in the first revision of audit sampling standards in Germany in the past thirty years. Conceptual analyses show that audit sampling must be considered in any financial statement audit and that the presence of multiple methodological opportunities facilitates its broad-ranging application. However, the actual implementation of these procedures leads to potential flaws. In particular, a uniform utilization of audit sampling, suited to a variety of audit situations, is a goal that cannot be achieved by any of the currently available procedures provided through professional standards, whether national or international.

The second paper (Chapter 3) provides a conceptual framework for the prevention of major concerns identified in the first paper. The framework is abstracted from professional standards and matches statistical and non-statistical audit sampling approaches to specific auditing objectives, according to the objective that each procedure is suited to. The findings of the paper may help to prevent typical pitfalls by providing recommendations towards more defendable sampling routines for diverse accounting populations and error environments, potentially achieving more consistent decisions and thus higher audit quality. The derived concept also offers resolute and specific clarification of the individual sampling procedural requirements, allowing for an empirical investigation of the current use of audit sampling plans in the field.

This investigation is provided in the third paper (Chapter 4), which consists of an archival study on the use of statistical and non-statistical audit sampling methods. Our analysis of working papers detailing conducted audit applications indicates a coherence of client-specific properties as well as error characteristics and individual risk functions that were attributed to the engagements in respect of the choice of audit sampling procedures, audit scope and the handling of accounting errors. By empirically tracking the difficulties throughout both statistical and non-statistical sampling applications revealed within

Chapter 3 – namely sample size and the extent of error containment – we showed that inherent and control risk do not have a significant impact on audit scope or error projection. Moreover, the study demonstrated that audit experience and the availability of previous working papers seem to play a significant role in influencing an auditors' judgment, when in fact these factors should not influence the conversion of client- and error-specific information into audit sampling properties.

Based on the outcome of the latter empirical study, the fourth paper (Chapter 5) focuses on the possible impact of a revision in professional auditing standards on auditor behavior regarding the desired audit scope and a reasonable treatment of disclosed error causes. Regarding the intended revision, our analysis assumes a positive association between the new guidance and the future performance of sampling procedures. Combined with the empirical results from the third paper, the expectations involved point to a potential prevention of ongoing flawed auditor behavior by the acquisition of larger sample sizes and the offering of a better sensitization towards arbitrary error containment.

These expectations were tested further over two empirical studies. The fifth paper (Chapter 6) details an experimental assessment of sample size decisions among 179 professional auditors from various medium-sized to large auditing firms within the application of non-statistical audit sampling procedures in common audit settings. The study looks at the effect of an ISA 530-based semi-structured decision aid on auditor behavior, addressing the scope of substantive tests of details in an accounts receivable population. In accordance with anchoring and adjustment theory, the decision aid can be judged as generating significantly larger samples throughout most audit objectives. However, the participants made disastrously insufficient adjustments of sample size in critical high-risk auditing situations, even though they seemed to be generally aware of the factors that impact sample size. Contrary to expectations, this effect is even stronger in the presence of the examined aid, when in fact the aid should moderate the participants' flaws.

The sixth and last paper (Chapter 7) extends the view from the preceding experiment by addressing the auditor's choice not to project but rather to contain an error from the remainder of an examined population when an actual representative error is revealed in substantial testing. As the archival study in the third paper had already revealed non-projection decisions to be everyday practice, the follow-up experiment examined the effect of error containment guidance on the participants' perceived occurrence of anomalies in audit sampling. Hypotheses based on the representative heuristic were tested among

80 students. Again, and contrary to the normative intent of ISA 530, the additional guidance was found not to have a significant effect on auditor behavior when providing bias avoidance techniques. In fact, the results indicate that containment decisions are influenced mainly by error-specific properties such as error frequency, rather than by the actual cause of an error.

To summarize, this dissertation provides important insights for standard setters concerning auditor performance in the application of audit sampling. All empirical results highlight the need for additional and more concrete guidance. This is also an important finding for auditing firms, which requires a reaction from auditing policy departments, as inadequate audit scopes and arbitrary error containment eventually lead to indefensible audit opinions, and ultimately establish a potential for renewed accounting and auditing scandals. Since investigations of professional auditors' performance in non-statistical audit sampling have so far been widely underrepresented in auditing research, this thesis brings to light necessary reactions of practicing auditors to changes in professional auditing standards as well as the variety of auditing situations they are faced with in the present and will be in the future.

9 References

Abdolmohammadi, M., and A. Wright. 1987. An examination of the effects of experience and task complexity on audit judgments. *The Accounting Review* 62 (1): 1–13.

Akresh, A., and K. Tatum. 1988. Audit sampling – dealing with the problems. *Journal of Accountancy* 166 (6): 58–64.

Allen, R., and R. Elder. 2005. A Longitudinal Investigation of Auditor Error Projection Decisions. *Auditing: A Journal of Practice and Theory* 24 (2): 69–84.

Arens, A., and J. Loebbecke. 1999. *Auditing: An integrated Approach*. 8th edition. Englewood Cliffs, NJ: Prentice Hall.

American Institute of Certified Public Accountants (AICPA) Auditing Standards Board. 2011. *Audit Sampling*. AU-C Section 530. New York, NY: AICPA.

American Institute of Certified Public Accountants (AICPA) Audit Sampling Committee. 2014. *Audit Sampling: Audit Guide*. New York, NY: AICPA.

American Institute of Certified Public Accountants (AICPA) Audit Sampling Committee. 2017. *Audit Sampling: Audit Guide*. New York, NY: AICPA.

American Institute of Certified Public Accountants (AICPA) Statistical Sampling Subcommittee. 1983. *Audit Sampling*. Audit and Accounting Guide. New York, NY: AICPA.

Ashton, R. 1983. *Research in Audit Decision Making: Rationale, Evidence, and Implications*. Research monography no. 6. Vancouver: Canadian Certified General Accountants' Research Foundation.

Ashton, R. 1990. Pressure and Performance in Accounting Decision Settings: Paradoxical Effects of Incentives, Feedback, and Justification. *Journal of Accounting Research* 78 (Supplement): 148–180.

Ashton, R., and J. Willingham. 1988. *Using and evaluating audit decision aids*. Audit Symposium IX: Proceedings of the 1988 Touche Ross/University of Kansas

Symposium on Auditing Procedures, edited by R. P. Srivastava and J. E. Rebele. University of Kansas.

Baetge, J. 1986. Auswahlprüfungen auf der Basis der Systemprüfung. In H. Albach, H. Albert, and H. Angermann (Ed.), *Wirtschaft und Wissenschaft im Wandel. Festschrift für Carl Zimmerer zum 60. Geburtstag* (pp. 45–63). Frankfurt am Main: Knapp.

Barnett, A., and W. Read. 1986. Sampling in Small Business Audits - How are Auditors implementing SAS No. 39? *Journal of Accountancy* 66 (1): 78–88.

Baumeister, D., C. Oldewurtel. 2016a. Stichproben in der Jahresabschlussprüfung - Voraussetzungen zur Anwendung anerkannter Erhebungsverfahren. *WP Praxis* 5 (7): 169–175.

Baumeister, D., C. Oldewurtel. 2016b. Die praktische Durchführung von Auswahlprüfungen - Strategien zur Vermeidung wesentlicher Fallstricke bei der Erhebung und Auswertung von Stichproben. *WP Praxis* 5 (8): 199–204.

Baumeister, D., C. Oldewurtel, C. Pott, and M. Weinand. 2018. Prüfungsnachweise mit Hilfe von Auswahlprüfungen - Empirische Analyse unter Berücksichtigung von IDW PS 300 n. F. und IDW PS 310. *Die Wirtschaftsprüfung* 71 (2): 73–82.

Bedard, J., and S. Biggs. 1981. The effect of domain-specific experience on evaluation of management representations in analytical procedures. *Auditing: A Journal of Practice and Theory* 10 (Supplement): 77–90.

Bedard, J, and K. Johnstone. 2010. Audit partner tenure and audit planning and pricing. *Auditing: A Journal of Practice and Theory* 29 (2): 45–70.

Bierkämper, M., and M. Toll. 2015. Einsatz analytischer Prüfungshandlungen im Rahmen der Jahresabschlussprüfung - Eine empirische Untersuchung. *WP Praxis* 4 (10): 248–255.

Blocher, E., and J. Bylinski. 1985. The influence of sample characteristics in sample evaluation. *Auditing: A Journal of Practice and Theory* 5 (1): 79–90.

Boatsman, J., C. Moeckel, and B. Pei. 1997. The effects of decision consequences on auditors' reliance on decision aids in audit planning. *Organizational Behavior and Human Decision Processes* 71 (2): 211–247.

Bonner, S., R. Libby, and M. Nelson. 1996. Using decision aids to improve auditors' conditional probability judgments. *The Accounting Review* 71 (2): 221–240.

Blocher, E., and J. Bylinski. 1985. The influence of sample characteristics in sample evaluation. *Auditing: A Journal of Practice and Theory* 5 (1): 79–90.

Burgstahler, D., and J. Jiambalvo. 1986. Sample error characteristics and projection of error to audit populations. *The Accounting Review* 61 (2): 233–248.

Burgstahler, D., S. Glover, and J. Jiambalvo. 2000. Error projection and uncertainty in the evaluation of aggregate error. *Auditing: A Journal of Practice and Theory* 19 (1): 79–99.

Butler, S. 1986. Anchoring in the judgmental evaluation of audit samples. *The Accounting Review* 61 (1): 101–111.

Christensen, B., R. Elder, and S. Glover. 2015. Behind the Numbers: Insights into Large Audit Firm Sampling Policies. *Accounting Horizons* 29 (1): 61–82.

Cohen, J., and T. Kida. 1989. The impact of analytical review results, internal control reliability, and experience on auditors' use of analytical review. *Journal of Accounting Research* 27 (2): 263–276.

Colbert, J. 1991. Statistical or Non-Statistical Sampling: Which Approach Is Best? *The Journal of Applied Business Research* 7 (2): 117–120.

Deutscher Steuerberaterverband e.V., AK Rechnungslegung. 2015. *Stellungnahme B 11/15*. Retrieved September 10, 2018, from https://www.idw.de/blob/86276/dc89972e5189794987616f7baa9b1797/down-idweps310-dstv-data.pdf.

Dubé-Rioux, L. and E. Russo. 1988. An availability bias in professional judgment. *Journal of Behavioral Decision Making* 1 (4): 223–237.

Durney, M., R. Elder, and S. Glover. 2014. Error rates, error projection, and consideration of sampling risk: Audit sampling data from the field. *Auditing: A Journal of Practice and Theory* 33 (2): 79–110.

Dusenbury, R., J. Reimers, and S. Wheeler. 1994. The effect of containment information and error frequency on projection of sample errors to audit populations. *The Accounting Review* 69 (1): 257–264.

Elder, R., and R. Allen. 1998. An empirical investigation of the auditors' decision to project errors. *Auditing: A Journal of Practice and Theory* 17 (2): 71-87.

Elder, R., and R. Allen. 2003. A longitudinal field investigation of auditor risk assessments and simple size decisions. *The Accounting Review* 78 (4): 983–1002.

Elder, R., A, Akresh, S. Glover, J. Higgs, and J. Liljegren. 2013. Audit sampling research: A synthesis and implications for future research. *Auditing: A Journal of Practice and Theory* 32 (Supplement 1): 99–129.

Elliott, R. 1983. Unique audit methods: Peat Marwick International. *Auditing: A Journal of Practice and Theory* 2 (2): 1–19.

Fay, R., J. Jenkins, and V. Popova. 2015. Effects of awareness of prior-year testing strategies and engagement risk on audit decision. *Managerial Auditing Journal* 30 (3): 226–243.

Fienberg, S., J. Neter, and R. Leitch. 1977. Estimating the Total Overstatement Error in Accounting Populations. *Journal of the American Statistical Association* 72 (358): 295–302.

Fischhoff, B. 1982. Debiasing. In Kahneman, D., P. Slovic; and A. Tversky, *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Freitas, A., P. Gollwitzer, and Y. Trope. 2004. The influence of abstract and concrete mindsets on anticipating and guiding others' self-regulatory efforts. *Journal of Experimental Social Psychology* 40 (6): 739–752.

Giezek, B. 2011. *Monetary Unit Sampling*. Wiesbaden: Springer.

Giezek, B. 2014. Monetary Unit Sampling in der Praxis - Ein empirischer Vergleich zum Einsatz von Stichprobenverfahren in den Jahren 1994, 2009 und 2013. *Die Wirtschaftsprüfung* 67 (11): 564–569.

Glass, A., K. Holyoak, and J. Santa. 1979. *Cognition*. Boston, MS: Addison-Wesley Publishing.

Göb, R, and M. Karrer. 2010. Die neue Aktualität der statistischen Stichprobenprüfung. *Die Wirtschaftsprüfung* 63 (11): 593–602.

Graham, L., J. Bedard, and D. Saurav. 2018. Managing Group Audit Risk in a Multi-component Audit Setting. *International Journal of Auditing* 22 (1): 40–54.

Graumann, M. 2015. *Wirtschaftliches Prüfungswesen*. Herne: NWB.

Griffin, P., and A. Wright. 2015. Commentaries on Big Data's Importance for Accounting and Auditing. *Accounting Horizons* 29 (2): 377–379.

Guy, D., D. Carmichael, and O. Whittington. 2002. *Audit Sampling: An Introduction*. New York, NY: Wiley.

Hall, T., J. Hunton, E. James, and B. Pierce. 2002. Sampling Practices of Auditors in Public Accounting, Industry, and Government. *Accounting Horizons* 16 (2): 125–136.

Hall, T., A. Higson, B. Pierce, K. Price, and C. Skousen. 2012. Haphazard sampling: Selection biases induce by control listing properties and the estimation consequences of theses biases. *Behavioral Research in Accounting* 24 (2): 101–132.

Ham, J., D. Losell, and W. Smieliauskas. 1987. Some empirical evidence on the stability of accounting error characteristics over time. *Contemporary Accounting Research* 4 (1), 210–226.

Hansen, J., and M. Waenke. 2010. Truth From Language and Truth From Fit: The Impact of Linguistic Concreteness and Level of Construal on Subjective Truth. *Personality and Social Psychology Bulletin* 36 (11): 1576–1588.

Heese, K., and M. Braatsch. 2013. Praktische Anwendung der ISA in Deutschland - Stichprobenprüfung (ISA 530). *Die Wirtschaftsprüfung* 66 (17): 841–848.

Hermanson, H. 1997. The effects of audit structure and experience on auditors' decisions to isolate errors. *Behavioral Research in Accounting* (Supplement): 76–93.

Hill, T. 1988. Random-Number Guessing and the First Digit Phenomenon. *Psychological Report* 62 (5): 967–971.

Hitzig, N. 1995. Audit sampling: a survey of current practice. *The CPA Journal* 65 (7): 54–57.

Hitzig, N. 2001. The mythical isolated error. *The CPA Journal* 71 (9): 50.

Hitzig, N. 2004. Statistical Sampling Revisited. *The CPA Journal* 74 (5): 30–35.

Hömberg, R. 1997. Zur Anwendung statistischer Prüfungsmethoden in der Wirtschaftsprüfung. *Betriebswirtschaftliche Forschung und Praxis* 49 (3): 245–265.

Hoogduin, L., T. Hall, and J. Tsay. 2010. Modified Sieve Sampling: A Method for Single- and Multi-Stage Probability- Proportional-to-Size Sampling.

Houghton, C., and J. Fogarty. 1991. Inherent risk. *Auditing: A Journal of Practice and Theory* 10 (1): 1–21. *Auditing: A Journal of Practice and Theory* 2 (2): 125–148.

Institute of Public Auditors in Germany (IDW). 1988. *Zur Anwendung stichprobengestützter Prüfungsmethoden bei der Jahresabschlussprüfung*. IDW Stellungnahme HFA 1/1988. *Die Wirtschaftsprüfung* (41): 240–247.

Institute of Public Auditors in Germany (IDW). 2001. *Analytische Prüfungshandlungen*. IDW Prüfungsstandard PS 312 Düsseldorf: IDW.

Institute of Public Auditors in Germany (IDW). 2011. *Einsatz von Datenanalysen im Rahmen der Abschlussprüfung*. IDW Prüfungshinweis PH 9.330.3 Düsseldorf: IDW.

Institute of Public Auditors in Germany (IDW). 2012a. *WP-Handbuch 2012, Wirtschaftsprüfung, Rechnungslegung, Beratung*. Düsseldorf: IDW.

Institute of Public Auditors in Germany (IDW). 2012b. *Feststellung und Beurteilung von Fehlerrisiken und Reaktionen des Abschlussprüfers auf die beurteilten Fehlerrisiken*. IDW Prüfungsstandard PS 261 n.F. Düsseldorf: IDW.

Institute of Public Auditors in Germany (IDW). 2012c. *Wesentlichkeit im Rahmen der Abschlussprüfung*. IDW Prüfungsstandard PS 250 n.F. Düsseldorf: IDW.

Institute of Public Auditors in Germany (IDW). 2015. *Ziele und allgemeine Grundsätze der Durchführung von Abschlussprüfungen*. IDW Prüfungsstandard PS 250. Düsseldorf: IDW.

Institute of Public Auditors in Germany (IDW). 2016a. *Prüfungsnachweise im Rahmen der Abschlussprüfung*. IDW Prüfungsstandard PS 300 n.F. Düsseldorf: IDW.

Institute of Public Auditors in Germany (IDW). 2016b. *Repräsentative Auswahlverfahren (Stichproben) in der Abschlussprüfung*. IDW Prüfungsstandard PS 310. Düsseldorf: IDW.

Institute of Public Auditors in Germany (IDW). 2016c. *Fragen und Antworten: Zur Durchführung einer repräsentativen Auswahl (Stichprobe) nach ISA 530 bzw. IDW EPS 310 oder einer bewussten Auswahl nach ISA 500 bzw. IDW EPS 300 n.F*. F & A zu ISA 530 bzw. IDW EPS 310 oder ISA 500 bzw. IDW EPS 300 n.F. Düsseldorf: IDW.

International Auditing and Assurance Standards Board (IAASB). 2009a. *Audit Evidence*. International Standard on Auditing 500. New York, NY: IAASB.

International Auditing and Assurance Standards Board (IAASB). 2009b. *Audit Sampling*. International Standard on Auditing 530. New York, NY: IAASB.

Jacoby, J., and N. Hitzig. 2011. Auditing Internal Controls in Small Populations. *The CPA Journal* 80 (12): 34–36.

Jiambalvo, J., and W. Waller. 1984. Decomposition and assessments of audit risks. *Auditing: A Journal of Practice and Theory* 3 (2): 80–88.

Johnson, J., R. Leitch, and J. Neter. 1981. Characteristics of Errors in Accounts Receivable and Inventory Accounts. *The Accounting Review* 56 (2): 270–293.

Joyce, E. and G. Biddle. 1981. Anchoring and adjustment in probabilistic inference in auditing. *Journal of Accounting Research* 19 (1): 120–145.

Jung, M., and H. Kellerer. 1995. Heterograde Annahmestichprobe versus Monetary-Unit Sampling. *Österreichische Zeitschrift für Rechnungswesen* 5 (8): 248–251.

Kachelmeier, S., and W. Messier, Jr. 1990. An investigation of the influence of a non-statistical decision aid on auditor sample size decisions. *The Accounting Review* 65 (1): 209–226.

Kahneman, D., and A. Tversky. 1972. Subjective Probability: A judgment of representativeness. *Cognitive Psychology* 3 (3): 430–454.

Kahneman, D., and A. Tversky. 1973. On the psychology of prediction. *Psychological Review* 80 (4): 237–251.

Kinney, W., and W. Uecker. 1982. Mitigating the consequences of anchoring in auditing judgments. *The Accounting Review* 57 (1): 55–69.

Kleinmuntz, B. 1990. Why we still use our heads instead of formulas: toward an integrative approach. *Psychological Bulletin* 107 (3): 296–310.

Kowalczik, T., and E. Wolfe. 1998. Anchoring effects associated with recommendations from expert decision aids: an experimental analysis. *Behavioral Research in Accounting* 10: 147–170.

Krahel, J., and W. Titera. 2015. Consequences of Big Data and Formalization on Accounting and Auditing Standards. *Accounting Horizons* 29 (2): 409-422.

Krehl, H. 2007. Stichproben und statistische Verfahren im Prüfungsprozess – das notwendige aber ungeliebte Kind. In Deggendorfer Forum zur digitalen Datenanalyse e. V. (Ed.), *Prozessoptimierung mit digitaler Datenanalyse - Ansätze und Methoden* (pp. 113–128). Berlin: Erich Schmidt Verlag.

Küster, B., and I. Bernhard. 2015. Praktische Anwendung der ISA in Deutschland -Prüfungsnachweise (ISA 500) und besondere Überlegungen zu ausgewählten Sachverhalten (ISA 501). *Die Wirtschaftsprüfung* (68): 1212–1222.

Leslie, D., A. Teitlebaum, and R. Anderson. 1979. *Dollar Unit Sampling. A Practical Guide for Auditors*. Toronto: Copp Clark Pitman.

Libby, R. 1985. Availability and the generation of hypotheses in analytical review. *Journal of Accounting Research* 23 (2): 648–667.

Maingot, M., and T. Quon. 2009. Sampling practices of internal auditors at corporations in the Standard & Poor's Toronto stock exchange composite index. *Accounting Perspectives* 8 (3): 215–234.

Mauldin, E., and C. Wolfe. 2014. How do auditors address control deficiencies that bias accounting estimates? *Contemporary Accounting Research* 31 (3): 658–680.

Messier, W., Jr. 1983. The effect of experience and firm type on materiality/disclosure judgments. *Journal of Accounting Research* 21 (2): 611–618.

Messier, W., Jr. 1995. Research in and development of audit decision aids. In *Judgment and decision making in accounting and auditing*. edited by R. H. Ashton and A. H. Ashton. Cambridge: Cambridge University Press.

Messier, W., Jr., S. Kachelmeier, and K. Jensen. 2001. An experimental assessment of recent professional developments in non-statistical audit sampling guidance. *Auditing: A Journal of Practice and Theory* 20 (1): 81–96.

Mochty, L. 2003. Mustererkennende Stichprobenprüfung – Verbindung der bewussten und zufälligen Auswahl von Prüfobjekten. In R. Richter (Ed.), *Entwicklungen der Wirtschaftsprüfung: Prüfungsmethoden - Risiko - Vertrauen* (pp. 185–224). Berlin: Erich Schmidt Verlag.

Mochty, L. 2007. Nichtstatistische Stichproben. In H.-J. Kirsch & S. Thiele (Ed.), *Rechnungslegung und Wirtschaftsprüfung. Festschrift zum 70. Geburtstag von Jörg Baetge* (pp. 1055–1083). Düsseldorf: IDW.

Mochty, L. 2012. Stichprobentechnik für Statistik-averse Wirtschaftsprüfer. In H. Schröder, V. Clausen, and A. Behr (Ed.), *Essener Beiträge zur empirischen Wirtschaftsforschung. Festschrift für Prof. Dr. Walter Assenmacher*. (pp. 75–106). Wiesbaden: Springer Gabler.

Mock, T., and J. Turner. 1981. *Internal Accounting Control Evaluation and Auditor Judgment*. Audit Research Monograph No. 3. New York, NY: AICPA.

Mock, T., and A. Wright. 1999. Are Audit Program Plans Risk-Adjusted? *Auditing: A Journal of Practice and Theory* 18 (1): 55-74.

Nanni, A. 1984. An exploration of the mediating effects of auditor experience and position in internal accounting control evaluation. *Accounting, Organizations and Society* 9 (2): 149–163.

Newiak, M. 2009. *Prüfungsurteile mit Dollar Unit Sampling. Ein Vergleich von Fehlerschätzmethoden für Zwecke der Wirtschaftsprüfung; Praxis, Theorie, Simulation*. Potsdam: Lehrstuhl für Statistik und Ökonometrie, Wirtschafts- und Sozialwissenschaftliche Fakultät der Universität Potsdam.

Nisbett, R., and L. Ross. 1980. *Human inference: Strategies and shortcomings in social judgment*. Upper Saddle River, NJ: Prentice Hall.

Peek, L., J. Neter, and C. Warren. 1991. AICPA nonstatistical audit sampling guidelines: A simulation. *Auditing: A Journal of Practice and Theory* 10 (2): 33–48.

Public Company Accounting Oversight Board (PCAOB). 2008. *Report on the PCAOB's 2004, 2005, 2006, and 2007 Inspections of Domestic Annually Inspected Firms*. PCAOB Release No. 2008-008. December 5. Washington, DC: PCAOB.

Public Company Accounting Oversight Board (PCAOB). 2014. Matters Related to Auditing Revenue in an Audit of Financial Statements. *Staff Audit Practice Alert No. 12*. New York, NY: PCAOB.

Rohrmann, B. 1986. Evaluating the usefulness of decision aids: A methodological perspective. *New Directions in Research on Decision Making*. Edited by B. Brehmer, H. Jungermann, P. Lourens, and G. Sevon, 363-382. Amsterdam: North-Holland.

Ruhnke, K., and A. von Torklus. 2008. Monetary Unit Sampling - Eine Analyse empirischer Studien. *Die Wirtschaftsprüfung* 61 (23): 1119–1128.

Schooler, J., D. Gerhard, and E. Loftus. 1986. Qualities of the unreal. *Journal of Experimental Psychology* 12 (2): 171–181.

Schwartz, D. 1997. Audit Sampling - A Practical Approach. *The CPA Journal* 67 (2): 56–59.

Sibelman, H. 2014. Myths and inconvenient truths about audit sampling. *The CPA Journal*. 84 (4): 6–10.

Smith, J., and T. Kida. 1991. Heuristics and biases: Expertise and task realism in Auditing. *Psychological Bulletin* 109 (3): 472–489.

Stael von Holstein, C. 1972. Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior & Human Performance* 8 (1): 139-158.

Stewart, T. 2012. *Technical Notes on the AICPA Audit Guide Audit Sampling*. New York, NY: AICPA.

Swearingen, J., and D. Hansen. 1990. Statistical and Non-Statistical Samples: Some Empirical Evidence. *The Journal of Applied Business Research* 6 (4): 49–58.

Taylor, T., and M. Dunnette. 1974. Relative contribution of decision-maker attributes to decision processes. *Organizational Behavior and Human Performance* 12 (2): 286–298.

Tversky, A., and D. Kahneman. 1971. Belief in the law of small numbers. *Psychological Bulletin* 76 (2): 105–110.

Tversky, A., and D. Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185 (4157): 1124–1131.

Tversky, A. 1977. Features of similarity. *Psychological Review* 84 (4).

Trompeter, G., and A. Wright. 2010. The world has changed—have analytical procedure practices? *Contemporary Accounting Research* 27 (2): 669–700.

Uecker, W., and W. Kinney. 1977. Judgmental evaluation of sample results: A study of the type and severity of errors made by practicing CPAs. *Accounting, Organizations and Society* 2 (3): 269–275.

Walo, J. 1995. The effects of client characteristics on audit scope. *Auditing: A Journal of Practice and Theory* 14 (1): 115–124.

Weinand, M., and M. Wolz. 2012. Der Einsatz von Auswahlverfahren im Rahmen der handelsrechtlichen Jahresabschlussprüfung - Techniken, Risiken und zu vermeidende Nebenwirkungen. *WP Praxis* 1 (4): 68–74.

Weinand, M., and M. Wolz. 2013. Verfahren der Stichprobenprüfung im Rahmen der handelsrechtlichen Jahresabschlussprüfung - Darstellung am Beispiel des Monetary Unit Sampling. *WP Praxis* 2 (1): 13–17.

Wheeler, S., R. Dusenbury, and J. Reimers. 1997. Projecting sample misstatements to audit populations: Theoretical, professional and empirical considerations. *Decision Sciences* 28 (2): 261–268.

Winkler, R. 1967a. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association* 62 (319): 776–800.

Winkler, R. 1967b. The Quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association* 62 (320): 1105–1120.

Wirtschaftsprüferkammer, WPK (Hrsg.). 2015. *Zeitpunkt der verpflichtenden Anwendung der ISA in Deutschland*. Retrieved September 10, 2018, from http://www.wpk.de/neu-auf-wpkde/alle/2015/sv/zeitpunkt-der-verpflichtenden-anwendung-der-isa-in-deutschland/

Wirtschaftsprüferkammer, WPK (Hrsg.). 2016. *Satzung der Wirtschaftsprüferkammer über die Rechte und Pflichten bei der Ausübung der Berufe des Wirtschaftsprüfers und des vereidigten Buchprüfers (Berufssatzung für Wirtschaftsprüfer/vereidigte Buchprüfer – BS WP/vBP) vom 21. Juni 2016 (BAnz AT 22.07.2016 B1)*. Retrieved September 10, 2018, from https://www.wpk.de/fileadmin/documents/WPK/Rechtsvorschriften/BS-WPvBP.pdf

Wolz, M. 2003. *Wesentlichkeit im Rahmen der Jahresabschlussprüfung. Bestandsaufnahme und Konzeptionen zur Umsetzung des Materialitygrundsatzes*. Düsseldorf: IDW.

Wolz, M. 2004. Dollar Unit Sampling - ein modifiziertes Verfahren zur Beurteilung über- und unter-bewerteter Prüffelder. *Betriebswirtschaftliche Forschung und Praxis* 56 (1): 60–80.

Wysocki, K. 2002. Prüfungstheorie, messtheoretischer Ansatz. In W. Ballwieser, A. Coenenberg, and K. Wysocki (Ed.), *Handwörterbuch der Rechnungslegung und Prüfung* (pp. 1707–1715). Stuttgart: Schäffer-Poeschel.

Zaheen, K., J. Shabbir, and S. Gupta. 2013. A New Sampling Design for Systematic Sampling. *Communication in Statistics - Theory and Methods* 42 (18): 3359–3370.