Herausforderungen bei der Implementation computerisierter adaptiver Hochschulklausuren

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)

Vorgelegt dem Fachbereich 05

Psychologie und Sportwissenschaften

Der Goethe-Universität Frankfurt am Main

Von Aron Fink

Geboren am 31.Dezember 1991 in Greiz

Frankfurt am Main

Mai 2022

Dekanin:

Prof. Dr. Sonja Rohrmann

Gutachter:

Prof. Dr. Andreas Frey

Prof. Dr. Johannes Hartig

Datum der Disputation:

18.01.2023

Zusammenfassung

Im Rahmen der fortschreitenden Digitalisierung der Hochschullehre finden auch verstärkt elektronische Prüfungsformate Eingang in den Alltag von Hochschulen. Insbesondere elektronische Abschlussklausuren (E-Klausuren) bieten hier die Möglichkeit, die Prüfungsbelastung Hochschulehrender durch die Automatisierung weiter Teile der Klausurkonstruktion, -administration und -auswertung zu reduzieren. Die Integration digitaler Technologien in die Prüfungspraxis deutscher Hochschulen ermöglicht dabei nicht nur eine ökonomische Klausurkonstruktion, realitätsnähere Klausuren (z. B. durch die Nutzung fachspezifischer Standardsoftware), und den Einsatz innovativer Testbausteine (z. B. Integration von Multimediadateien in Items), sondern auch die Nutzung aktueller psychometrischer Methoden. Insbesondere die Konstruktion von Hochschulklausuren als kriteriumsorientierte, adaptive Tests (z. B. Spoden & Frey, 2021), hat das Potential Hochschulklausuren individualisierter, messpräzisier und fairer zu machen, sowie die Validität der aus der Klausurbearbeitung abgeleiteten Testwertinterpretationen zu steigern. Um kriteriumsorientierte, adaptive Hochschulklausuren in der Breite nutzbar zu machen, müssen allerdings zuvor einige Herausforderungen gemeistert werden, denen sich diese Arbeit widmet. Die in den vier Einzelarbeiten dieser Dissertation betrachteten Herausforderungen lassen sich auf einer psychometrischen, einer personalen und einer technischen Ebene verorten.

Auf der psychometrischen Ebene ist eine zentrale Herausforderung die ökonomische Kalibrierung des Itempools. Üblicherweise wird bei der Konstruktion adaptiver Tests eine dreistellige Anzahl an Items konstruiert und mittels einer separaten Kalibrierungsstudie im Vorlauf der operationalen Testanwendung mit mehreren hundert Testpersonen kalibriert. Die massierte Konstruktion vieler Items und die Durchführung einer zusätzlichen empirischen Studie lässt sich im Rahmen von Hochschulklausuren nur schwer realisieren. Im *ersten Einzelbeitrag* wird daher eine neuartige kontinuierliche Kalibrierungsstrategie (KKS) vorgestellt und im Rahmen einer Monte-Carlo-Simulation hinsichtlich ihrer psychometrischen Eigenschaften geprüft. Zusammenfassend ermöglicht die KKS, adaptive Tests während wiederkehrender Testanwendungen bei konstanter Berichtsmetrik, Kontrolle von Itemparameter-Drift und fortlaufender Ergänzung des Itempools zu kalibrieren. Es zeigt sich, dass die KKS selbst für sehr kleine Stichproben eine geeignete Methode darstellt, den Itempool über mehrere Testanwendungen hinweg fortlaufend zu kalibrieren.

Um die Berichtsmetrik dabei über die verschiedenen Testanwendungen hinweg konstant zu halten, und somit Vergleichbarkeit der Ergebnisse verschiedener Testzeitpunkte (z. B. Semester) zu gewährleisten, nutzt die KKS Equating-Methoden (z. B. Kolen & Brennan, 2014) zum Herstellen einer statistischen Verbindung zwischen Klausurdurchläufen. Die Qualität dieser

statistischen Verbindung hängt dabei von verschiedenen Parametern ab. Im *zweiten Einzelbeitrag* werden daher verschiedene Konfigurationen der in die KKS implementierten Equating-Prozedur hinsichtlich ihres Einflusses auf die Qualität der Parameterschätzungen im Rahmen einer Monte-Carlo-Simulation untersucht und auf Basis der Ergebnisse praktische Empfehlungen abgeleitet. Hierfür werden unter anderem die Schwierigkeitsverteilung der genutzten Linkitems sowie die verwendete Skalentransformationsmethode variiert. Es zeigt sich, dass die KKS unter verschiedenen Konfigurationen in der Lage ist, die Skala über mehrere Testzyklen hinweg konstant zu halten. Normal- beziehungsweise gleichverteile Schwierigkeitsverteilungen der Linkitems sowie die Stocking-Lord-Skalentransformationsmethode (Stocking & Lord, 1983) erweisen sich hierbei am vorteilhaftesten.

Auf personaler Ebene stellt die Akzeptanz seitens der Hochschullehrenden einen kritischen Erfolgsfaktor für die Implementation neuer E-Learning Systeme in Lehrveranstaltungen dar. Angelehnt an Technologieakzeptanzmodellen (z. B. Technology Acceptance Model; Davis, 1989) wird im *dritten Einzelbeitrag* ein empirisch prüfbares Modell – das Technology-based Exams Acceptance Model (TEAM) – zur Vorhersage der Intention zur Nutzung von adaptiven und nicht-adaptiven E-Klausursystemen seitens Hochschullehrender vorgeschlagen und anhand der Daten von N = 993 deutschen Hochschullehrenden empirisch geprüft. Das postulierte Modell weist einen guten Modellfit auf. Die Ergebnisse weisen die wahrgenommene Nützlichkeit als Schlüsselprädiktor für die Nutzungsintention aus. Medienbezogene Variablen haben indirekte Effekte auf die wahrgenommene Nützlichkeit, mediiert über vorherige Nutzungserfahrungen mit Bildungstechnologien. Darüber hinaus spielt die subjektive Norm eine wichtige Rolle bei der Erklärung der Akzeptanz von E-Klausuren. Das Modell bietet eine solide theoretische Grundlage, die für die erfolgreiche Einführung von adaptiven und nicht-adaptiven E-Klausuren im Hochschulbereich genutzt werden kann.

Auf der technischen Ebene stellt schließlich das Bereitstellen einer geeigneten Klausurensoftware zur Umsetzung kriteriumsorientierter, adaptiver Hochschulklausuren eine der größten Herausforderungen für deren Implementation dar. Im *vierten Einzelbeitrag* wird daher die im Rahmen der hier vorliegenden Dissertation entwickelte KAT-HS-App vorgestellt. Mit dieser können kriteriumsorientierte, adaptive Hochschulklausuren benutzerfreundlich konstruiert, administriert und analysiert werden. Darüber hinaus ist mit ihr auch die Anwendung der in den Einzelbeiträgen 1 und 2 beschriebenen KKS möglich.

In der Zusammenschau liegen mit dieser Dissertation also direkt einsetzbare Bausteine vor, die ein Meistern zentraler Herausforderungen bei der zielgerichteten Implementation kriteriumsorientierter, adaptiver E-Klausuren auf psychometrischer, personaler sowie technischer Ebene ermöglichen.

Abstract

In the context of the ongoing digitization of higher education, digital assessments are increasingly becoming part of teaching practices at universities. In particular, electronic high-stakes examinations (e-exams) offer the possibility to considerably reduce the examination load of university teachers, especially due to the time-saving effect of the automated coding of student responses. The integration of digital technologies into the examination practice of universities enables not only resource-oriented test development, more realistic exams (e.g., through the use of subject-specific standard software), and the use of innovative test modules (e.g., integration of multimedia files into items), but also the use of state-of-the-art psychometrical methods. Especially, the construction of university exams as criterion-referenced adaptive tests (e.g., Spoden & Frey, 2021) has the potential to make university exams more individualized, more precise, and fairer, as well as to increase the validity of the test score interpretations derived from student performance. In order to make use of criterion-referenced adaptive exams, however, some challenges first have to be overcome. The challenges considered in the four articles of this dissertation can be divided into a psychometric, personal, and technical level.

At the psychometric level, a key challenge is connected to the calibration of the item pool. Usually, for adaptive tests, several hundreds of items are constructed and are calibrated by means of a separate calibration study with several hundred test persons prior to the operational test phase. In the context of university exams, it is difficult to conduct a separate large calibration study and to construct a very large number of items for the calibration study. Therefore, in the *first article,* a novel continuous calibration strategy (CCS) is presented and investigated by means of a Monte Carlo simulation. In short, the CCS enables a step-by-step calibration of the item pool across several test applications without a separate calibration study, while maintaining the scale, controlling for item parameter drift, and continuously replenishing the item pool. It is shown that even for very small samples, the CCS is a suitable method to calibrate the item pool continuously across several test applications. Across test applications, the tests become increasingly more precise, deficient items are identified, and the scale is maintained.

In order to keep the scale constant across test applications and, thus, to ensure the comparability of results across different time points (e.g., semesters), the CCS uses equating methods (e.g., Kolen & Brennan, 2014) to establish a statistical link. The quality of this statistical link depends on various parameters. In the *second article*, different configurations of the equating procedure implemented in the CCS were investigated with regard to their influence on the quality of the parameter estimates in a Monte Carlo simulation. Based on the results,

practical recommendations were derived. For this purpose, among others, the difficulty distribution of the link items and the scale transformation method used were varied. Normal or uniform distributed difficulty distributions of the link items as well as the Stocking-Lord scale transformation method (Stocking & Lord, 1983) proved to be the most advantageous.

At the personal level, the acceptance by the teaching staff is a critical factor for the successful implementation of e-learning technologies in courses. In the *third article*, an empirically testable model—the Technology-Based Exams Acceptance Model (TEAM)—for predicting higher education teachers' intention to use adaptive and nonadaptive e-exams is proposed and tested empirically using data from $N = 993$ higher education teachers in Germany. It draws from existing technology acceptance models (e.g., Technology Acceptance Model; Davis, 1989). The model showed an acceptable fit. The results highlight the importance of perceived usefulness as the key predictor of the intention to use. Media-related variables had indirect effects on the perceived usefulness, which were mediated by prior experience with e-learning technologies for teaching purposes. Furthermore, the subjective norm played an important role in explaining e-exam acceptance. The model provides a solid theoretical basis that can be used for the successful implementation of adaptive and nonadaptive e-exams in higher education.

Finally, at the technical level, one of the biggest challenges for the implementation of criterion-referenced, adaptive university exams is the provision of a suitable e-exam software. The *fourth article* therefore presents the KAT-HS-App, an e-exam software developed in the context of this dissertation. The KAT-HS-App enables the construction, administration, and evaluation of criterion-referenced, adaptive university exams in a user-friendly manner. Furthermore, it enables users to apply the CCS described in the first and second articles of this dissertation.

In summary, this dissertation provides directly applicable building blocks that allow central challenges in the implementation of criterion-referenced, adaptive e-exams to be overcome on a psychometric, personal, and technical level.

Inhalt

1 Einleitung

Vor dem Hintergrund der vielfältigen Einflüsse der Digitalisierung auf die Hochschulbildung nimmt auch die Nutzung unterschiedlicher Formen elektronischer Assessments (E-Assessments) an deutschen Hochschulen zu (Bandtel et al., 2021). Der Begriff E-Assessment umfasst dabei aus prüfungsdidaktischer Sicht alle hauptsächlich kognitiven Leistungsmessungen, deren Durchführung und Verarbeitung mithilfe digitaler Informations- und Kommunikationstechnologien realisiert werden (Schmees & Horn, 2014). Ihr Einsatzgebiet im Hochschulbereich reicht von formativen Leistungsmessungen, wie beispielsweise semesterbegleitenden Selbsttests auf Lehrplattformen oder Quizzes, die in der Lehrveranstaltung über Audio-Response-Systeme durchgeführt werden, bis zu summativen Leistungsmessungen wie elektronischen Abschlussklausuren (im Weiteren: E-Klausuren). Unter letzteren versteht man Szenarien, „[…] bei denen Studierende in einem realen Raum zu einem festen Termin zusammenkommen und ihre Leistungen oder Arbeitsproben in einem elektronischen System zwecks Bewertung eingeben." (Häfer & Matthé, 2016, S. 195). Besonders in diesem Bereich der E-Assessments liegt für Lehrende die Möglichkeit, die Prüfungsbelastung erheblich zu reduzieren, da große Teile der Klausurdurchführung, -auswertung, und Ergebnisrückmeldung automatisiert werden können. In einer digitalisierten Welt sollte die Lehre an Hochschulen zudem ohnehin anhand ganzheitlicher E-Learning-Konzepte erfolgen, deren didaktisches Ziel es ist, den Studierenden Kompetenzen anschaulich zu vermitteln und sie in Prüfungssituationen mit den gleichen Werkzeugen arbeiten zu lassen, die sie auch während des Semesters genutzt haben (Schulz, 2016). Sie lernen anhand computergestützter Vorlesungsfolien, Lehrvideos oder Simulationen, die über Learning-Management-Systeme zur Verfügung gestellt werden, testen ihr Wissen über elektronische Selbsttests, nutzen über das Semester verschiedenste fachrelevante Software zur Bearbeitung von Aufgaben in Lehrveranstaltungen und schreiben schon lange ihre Semesterarbeiten am Computer. „Der mediale Bruch lauert für die Studierenden heutzutage in der schriftlichen Papierprüfung" (Schulz, 2016, S. 209). Für eine mündige gesellschaftliche Teilhabe in einer zunehmend technologisierten Welt stehen Hochschulen in der Verantwortung, nicht nur Lehr- und Lernkonzepte derart zu gestalten, dass Lernende mit den Mechanismen einer digitalen Gesellschaft kompetent, verantwortlich und kritisch umgehen können, sondern diese auch im Bereich des Prüfens aufzugreifen und Prüfungsformate entsprechend umzugestalten (Bedenlier et al., 2021).

Gleichzeitig ist im Zuge der Bologna-Reform insbesondere das Thema der Kompetenzorientierung in den Blickpunkt gerückt, wobei Kompetenzen hier im Sinne kontextspezifischer, kognitiver Leistungsdispositionen verstanden werden, die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen (Klieme & Leutner, 2006) – eine Entwicklung, die auch die Ausgestaltung von E-Klausuren an Hochschulen hochgradig beeinflusst.

Als abschließendes Element kompetenzorientierter Lehrveranstaltungen beziehungsweise Studienmodule sollten Klausuren nun im Sinne von Kompetenztests (z. B. Frey & Hartig, 2022) derart konzipiert werden, dass sie über die reine Abfrage gelehrter Inhalte hinaus die Beurteilung der individuellen Lernzielerreichung, das heißt vor allem die Einschätzung der individuellen Kompetenzniveaus der Studierenden, ermöglichen. Eine kompetenzorientierte Hochschullehre verlangt danach, die zentralen Elemente der Lehr-, Lern- und insbesondere auch der Prüfungsgestaltung von Anfang an konsequent auf die intendierten Lernziele zu beziehen. Dieser Ansatz wurde von Biggs (1996) als didaktisches Konzept des „Constructive Alignment“ beschrieben. Der Umstand, dass sich Studierende in der Ausgestaltung ihrer Lernprozesse häufig an den einer Lehrveranstaltung zugrundeliegenden Prüfungsanforderungen orientieren (Schaper & Hilkenmeier, 2013), macht Prüfungen zu wirkmächtigen didaktischen Werkzeugen zur Steuerung von Lernhandlungen der Studierenden (siehe z. B. Roßnagel et al., 2021 für empirische Effekte von Constructive Alignment auf lernbezogene Variablen im Hochschulbereich).

Aus testtheoretischer Sicht ist damit verbunden, dass Klausuren als kriteriumsorientierte Tests konstruiert werden sollten, also als Tests, deren „[…] Testwerte als kontinuierlich oder kategorial beschriebene Ausprägungen eines Individuums bezüglich einer wohldefinierten Inhalts- oder Verhaltensdomäne interpretiert werden können.“ (Herzberg & Frey, 2011, S. 283). Durch die Nutzung kriteriumsorientierter Testverfahren können von den individuellen Klausurergebnissen Rückschlüsse auf das Erreichen vorab definierter Kompetenzniveaus getroffen werden. Klausuren, ob papierbasiert oder elektronisch, bei denen dieselbe Menge an Items allen Studierenden präsentiert wird (sog. fixed item testing; FIT) treffen dabei aber auf ein Problem, welches in Anlehnung an das sogenannte Bandbreiten-Genauigkeits-Dilemma (engl.: bandwidth fidelity dilemma; BFD; Cronbach & Gleser, 1965) beschrieben werden kann. Im Bereich der psychologischen Diagnostik besagt das BFD, dass je breiter ein Test angelegt ist, desto ungenauer sind die resultierenden Messungen. Gegeben bestehender Ressourcen (z. B. Testzeit) muss daher bei jedem Test eine Abwägung getroffen werden, ob wenige Merkmale sehr präzise (also mit hoher Messgenauigkeit) oder ob viele Merkmale eher überblicksartig erfasst werden sollen (Ones & Viswesvaran, 1996). Diese Idee kann man anstelle der Messung mehrerer Merkmale, auf die in der Leistungs- beziehungsweise Kompetenzdiagnostik üblichen Verortung von Testpersonen auf einem latenten Merkmalskontinuum anwenden. Die Messgenauigkeit dieser Verortung ist dabei eng mit der Testinformation (i. S. d. Menge an diagnostischer Information, die ein Test in Bezug auf die Schätzung der individuellen Merkmalsausprägung enthält; siehe Kapitel 2.3.4) verknüpft. So haben die Ergebnisse eines bestimmten Tests an der Stelle die höchste Messgenauigkeit, an der dessen Informationsfunktion (i. S. d. Höhe der Testinformation in Abhängigkeit der individuellen Merkmalsauprägung) ihr Maximum erreicht (z. B. Embretson & Reise, 2000). Sind die Items eines Tests in etwa gleich schwierig, existiert ein sehr schmaler Bereich auf dem latenten Merkmalskontinuum, in dem eine hohe Messgenauigkeit erreicht wird. Dieser Bereich spiegelt die

Bandbreite wider. Das Dilemma besteht nun darin, dass gegeben konstanter Itemdiskrimination und Testlänge eine hohe Messgenauigkeit auf Kosten der Bandbreite erreicht wird und umgekehrt (McBride, 1976). Die Bandbreite ist damit eine direkte Funktion der Verteilung der Itemschwierigkeiten. Je mehr diese variieren, desto größer ist die Bandbreite; je weniger sie variieren, desto höher ist die Messgenauigkeit. Übertragen auf Hochschulklausuren bedeutet dies, dass traditionelle Klausuren, üblicherweise die höchste Präzision für Testpersonen mit mittlerer Kompetenzausprägung liefern, während die Präzision für Testpersonen mit extremen Testergebnissen, das heißt an den Rändern der Kompetenzverteilung, deutlich abnimmt (Dolan & Burling, 2017). Dies gefährdet den für Hochschulklausuren zentralen Grundsatz der Gleichbehandlung aller geprüften Studierenden, da die Genauigkeit der aus der Klausurbearbeitung abgeleiteten Bewertungen (häufig in Form von Noten oder als bestanden/nicht bestanden) abhängig von der individuellen Kompetenzausprägung der geprüften Person ist. Vor allem für Personen in den Randbereichen der Kompetenzverteilung ist die Genauigkeit der Bewertung aber häufig von besonderer Relevanz, da von diesen beispielsweise die Weiterführung des Studiums oder eine Studienförderung abhängt (Frey, 2021). Soll in einer Klausur die ganze Bandbreite der vorab definierten Kompetenzniveaus mit einer angemessenen Messpräzision geprüft werden, so sind sehr viele Items, verteilt über das gesamte Kompetenzspektrum von Nöten, was unter der Nutzung von FIT zu einer erheblichen Verlängerung der Testzeit führen würde.

Eine mögliche Lösung für dieses Problem stellt das computerisierte adaptive Testen (CAT; z. B. Frey, 2020) dar. Bei CAT orientiert sich die Itemauswahl während der Testung an der individuellen Merkmalsausprägung der Testperson. Der Testverlauf passt sich dabei dem gezeigten Antwortverhalten der Testperson an. Ziel dieses Vorgehens ist es, den Testpersonen nur solche Items vorzulegen, die in Bezug auf die individuelle Merkmalsausprägung möglichst viel diagnostische Information enthalten. Somit hängt die adaptive Itemauswahl zum einen von der individuellen Merkmalsausprägung und zum anderen von spezifischen psychometrischen Eigenschaften der Items ab. Das bedeutet, dass unter Verwendung von CAT im Idealfall jede Testperson Items mit adäquatem Schwierigkeitsniveau zur Bearbeitung vorgelegt bekommt, wodurch eine Angleichung der Messpräzision über das gesamte Kompetenzspektrum erreicht und somit das angesprochene Problem der mangelnden Differenzierungsfähigkeit klassischer Hochschulklausuren in den Randbereichen der Kompetenzverteilung gelöst werden kann. Zudem kann durch die Verwendung von CAT generell eine erhebliche Effizienzsteigerung im Sinne einer erhöhten Messgenauigkeit und/oder kürzeren Testlänge im Vergleich zu einer nicht-adaptiven Itemauswahl erzielt werden (Segall, 2005), was ihre Nützlichkeit für Hochschulklausuren weiter erhöht. Dieser Umstand kann auch als Grund dafür gesehen werden, dass CAT vor allem im englischsprachigen Raum bereits seit vielen Jahren erfolgreich bei großen Testprogrammen im high-stakes Bereich eingesetzt wird. Beispiele hierfür sind das National Council Licensure Examination (NCLEX; https://www.ncsbn.org/nclex.htm) oder der Graduate Management

Admission Test (GMAT; https://www.gmac.com/gmat-other-assessments; für eine Übersicht operationaler CATs siehe z. B. http://www.iacat.org/content/operational-cat-programs). Außerhalb von Forschung und groß angelegten Testprogrammen ist der Einsatz von CAT allerdings eher selten. Besonders im Bereich von Hochschulklausuren bietet CAT aber ein noch nicht genutztes Potential, Einschränkungen, die klassische Klausuren auf Basis von FIT mit sich bringen, zu überwinden und so die Qualität von Hochschulklausuren erheblich zu steigern (Spoden & Frey, 2021).

Um CAT auf den Bereich der Hochschulklausuren anzuwenden, müssen allerdings zuvor einige Herausforderungen gemeistert werden, denen sich diese Arbeit widmet. Die in dieser Arbeit betrachteten Herausforderungen lassen sich auf einer psychometrischen, personalen und einer technischen Ebene verorten.

Aus psychometrischer Sicht stellen vor allem die üblicherweise an Hochschulen vorzufindenden Rahmenbedingungen (z. B. vergleichsweise kleine Kalibrierungsstichproben, begrenzte Ressourcen für die Itementwicklung), eine Hürde für die Erstellung eines kalibrierten CAT-Itempools und somit für die Umsetzung von CAT im Hochschulbereich dar. Die gängige Prozedur, bei der in einer der operationalen CAT-Phase vorgelagerten Kalibrierungsstudie die zu kalibrierenden Items einer großen Anzahl an Personen zur Bearbeitung vorgelegt werden, ist im Routinebetrieb der meisten Hochschulen kaum realisierbar. Die erste Fragestellung dieser Dissertation lautet daher:

1. Wie kann die Kalibrierung eines Itempools für kriteriumsorientierte, adaptive Hochschulklausuren im laufenden Lehrbetrieb erfolgen?

Der Beantwortung von Fragestellung 1 widmet sich Beitrag 1 (siehe Kapitel 3.1). In diesem wird eine neuartige kontinuierliche Kalibrierungsstrategie (KKS) vorgestellt und ihre psychometrischen Eigenschaften im Rahmen einer Monte-Carlo-Simulation geprüft. Die KKS ermöglicht es, einen CAT-Itempool über mehrere Testanwendungen hinweg im laufenden Lehrbetrieb ohne separate Kalibrierungsstudie zu kalibrieren.

Neben der Testfairness innerhalb einer Kohorte (durch beispielsweise vergleichbare Messpräzision über das komplette Kompetenzspektrum), ist es für Hochschulklausuren zudem erstrebenswert, Vergleichbarkeit der Ergebnisse über verschiedene Testzeitpunkte hinweg (z. B. Semester) zu gewährleisten. Übliche Klausuren sind allerdings über Testzeitpunkte hinweg nicht statistisch verbunden, was dazu führen kann, dass trotz eines äquivalenten Kompetenzniveaus, die Klausurergebnisse zwischen Testzeitpunkten variieren können und somit nicht vergleichbar sind (Frey, Spoden & Born, 2020; Born & Fink, 2021). Innerhalb der in Beitrag 1 vorgestellten Kalibrierungsstrategie werden daher sogenannte Equating-Methoden (z. B. Kolen & Brennan, 2014) genutzt, um eine statistische Verbindung zwischen einzelnen Klausurdurchläufen herzustellen und die Ergebnisse verschiedener Klausurzeitpunkte somit direkt vergleichbar zu machen. Die Qualität

dieser statistischen Verbindung hängt von verschiedenen Parametern ab. Die zweite Fragestellung lautet daher:

2. Wie sollte die Equating-Prozedur der KKS konfiguriert sein, um bestmögliche Ergebnisse hinsichtlich der Genauigkeit der Parameterschätzungen sowie der Vergleichbarkeit der Klausurergebnisse über verschiedene Testzeitpunkte zu erzielen?

Beitrag 2 (siehe Kapitel 3.2) widmet sich der Beantwortung dieser Fragestellung, indem verschiedene Konfigurationen der in die KKS implementierten Equating-Prozedur hinsichtlich ihres Einflusses auf die Qualität der Parameterschätzungen im Rahmen einer Monte-Carlo-Simulation untersucht und auf Basis der Ergebnisse praktische Empfehlungen abgleitet werden.

Das Gelingen einer flächendeckenden Implementation neuartiger IT-Systeme hängt stark von der Akzeptanz solcher Systeme seitens der intendierten Nutzergruppen ab. Im Bereich von Hochschulklausuren sind dies vor allem die Lernenden sowie die Lehrenden. Bisher lag der Fokus der Studien zu dieser Thematik vor allem auf der Akzeptanz seitens der Lernenden (z. B. Maqableh et al., 2015; Terzis & Economides, 2011; Terzis et al., 2012; Zheng & Bender, 2019). Dabei ist aber vor allem die Akzeptanz seitens der Lehrenden ein kritischer Faktor für die erfolgreiche Implementation neuer E-Learning Systeme in Lehrveranstaltungen (Bennett et al., 2017; Brady et al., 2019; Nikou & Economides, 2018; Paiva et al., 2017). Um zeit- und ressourcenaufwändige Fehler im Implementationsprozess kriteriumsorientierter, adaptiver Hochschulklausuren zu vermeiden, liegt der Fokus auf personaler Ebene daher auf Faktoren, die die Intention zur Nutzung von adaptiven E-Klausursystemen seitens Hochschullehrender beeinflussen.

3. Welche Faktoren beeinflussen die Intention zur Nutzung von kriteriumsorientierten adaptiven Hochschulklausuren seitens Hochschullehrender?

Zur Beantwortung von Fragestellung 3 wird im dritten Beitrag (siehe Kapitel 3.3) ein empirisch prüfbares Modell zur Vorhersage der Intention zur Nutzung von adaptiven und nicht-adaptiven E-Klausursystemen seitens Hochschullehrender entwickelt und anhand empirischer Daten geprüft.

Schließlich stellt auf der technischen Ebene die Bereitstellung einer geeigneten Klausurensoftware eine der größten Herausforderungen für die Implementation kriteriumsorientierter adaptiver Hochschulklausuren dar. Für die Anwendung verschiedener Methoden der Item Response Theory (IRT; z. B. van der Linden, 2016b), die für die Umsetzung adaptiver Hochschulklausuren notwendig sind, sind umfassende psychometrische Kenntnisse und die Einarbeitung in verschiedene Softwarepakete nötig, über die viele potentielle Anwenderinnen und Anwender im Hochschulbereich nicht verfügen. Im Hinblick auf die Nachhaltigkeit eines E-Klausursystems ist es ferner erstrebenswert, eine dokumentierte, transparente und seitens der Hochschulen anpassbare Software bereitzustellen, die nicht in Abhängigkeitsverhältnissen und

Lizenzkosten mündet. Forschungsfrage 4 lautet daher:

4. Wie kann die Konstruktion, Administration und Auswertung kriteriumsorientierter adaptiver Hochschulklausuren softwaretechnisch auch für Personen mit geringeren psychometrischen Vorkenntnissen benutzerfreundlich ermöglicht werden?

Die im Rahmen des Dissertationsvorhabens entwickelte KAT-HS-App (KAT-HS = kriteriumsorientiertes adaptives Testen in der Hochschule) liefert die Antwort auf diese Frage. Ihre Kernfunktionalitäten werden in Beitrag 4 (siehe Kapitel 3.4) beschrieben. Die KAT-HS-App zielt darauf ab, die Konstruktion, Administration und Auswertung computerbasierter Tests auf dem aktuellen wissenschaftlichen Stand nicht nur für Expertinnen und Experten, sondern auch für Nutzergruppen mit geringeren psychometrischen Vorkenntnissen benutzerfreundlich zu ermöglichen. Zudem ermöglicht die App die in den Beiträgen 1 und 2 beschriebene KKS.

Der Rest der Arbeit gliedert sich wie folgt: Zunächst wird in die zentralen theoretischen Grundlagen und Begrifflichkeiten zu adaptiven Hochschulklausuren eingeführt (Kapitel 2). Den theoretischen Grundlagen schließt sich die Darstellung der Einzelbeiträge an (Kapitel 3). Schließlich folgt eine abschließende Gesamtdiskussion (Kapitel 4), in der die Fragestellungen beantwortet sowie der Erkenntnisgewinn und offengebliebene Fragen der Arbeit diskutiert werden.

2 Computerisiertes Adaptives Testen für Hochschulklausuren

In diesem Kapitel wird in die Thematik des adaptiven Testens eingeführt sowie zentrale Aspekte von CAT unter Bezugnahme auf Hochschulklausuren erläutert. Dazu werden in Kapitel 2.1 zunächst die Grundlagen des adaptiven Testens dargestellt. Kapitel 2.2 gibt eine kurze Einführung in die testtheoretische Grundlage von CAT – die IRT. In Kapitel 2.3 werden die elementaren Bausteine von CAT beschrieben sowie konkrete Gestaltungshinweise in Bezug auf typische Hochschulklausuren abgeleitet. In Kapitel 2.4 wird schließlich ein zusammenfassender Überblick der zentralen Vorteile des Einsatzes von CAT für Hochschulklausuren gegeben.

2.1 Grundlagen des computerisierten adaptiven Testens

Üblicherweise wird bei Testverfahren, so auch bei traditionellen Hochschulklausuren, jeder Testperson die gleiche vorab definierte Menge an Items in einer festen Reihenfolge zur Bearbeitung vorgelegt. Bei CAT hingegen orientiert sich die Itemauswahl am bisher im Testverlauf gezeigten Antwortverhalten der Testperson. Auch wenn CAT im Allgemeinen als innovativ angesehen wird, ist die Idee dahinter nicht neu. Orientiert man sich am Beispiel mündlicher Prüfungen, so ist jedem Prüfenden klar, dass es wenig Sinn macht, dem Prüfling wiederholt zu leichte beziehungsweise zu schwere Fragen zu stellen, da mit hoher Wahrscheinlichkeit vorausgesagt werden kann, ob die zu prüfende Person die Fragen richtig beziehungsweise falsch beantworten wird. Das wiederholte Stellen zu leichter oder zu schwerer Fragen liefert somit kaum neue diagnostische Information bezüglich der Merkmalsausprägung des Prüflings. In der Regel versucht der Prüfende sich stattdessen mit dem Schwierigkeitsgrad der Prüfungsfragen schrittweise an das Kompetenzniveau des Prüflings anzunähern, um so das tatsächliche Kompetenzniveau möglichst genau (d. h. mit möglichst geringem Messfehler) bestimmen zu können. Üblicherweise resultiert aus einem solchen Vorgehen, dass Personen mit einer hohen Merkmalsausprägung schwierigere Fragen gestellt bekommen als Personen mit einer niedrigeren Merkmalsausprägung. Wie erfolgreich ein solches Vorgehen tatsächlich ist, hängt im Wesentlichen von zwei Faktoren ab: a) der Fähigkeit des Prüfenden das Kompetenzniveau des Prüflings nach Beantwortung einer Frage einzuschätzen und b) der Kenntnis des jeweiligen Schwierigkeitsgrades der für die Prüfung vorbereiteten Fragen (van der Linden, 2018).

Das Vorgehen bei CAT ist hier vergleichbar, den individuellen Testverlauf bestimmt aber anstatt des Prüfenden der Computer. So bekommen auch hier Testpersonen nur diejenigen Items zur Bearbeitung vorgelegt, die besonders informativ bezüglich des zu messenden Merkmals sind. Wie bereits erwähnt ist dies dann der Fall, wenn der Schwierigkeitsgrad der Items mit der individuellen Merkmalsausprägung korrespondiert (van der Linden & Glass, 2010). Somit muss auch der Computer a) dazu befähigt sein, das Kompetenzniveau des Prüflings in Echtzeit einschätzen zu

können und b) Kenntnisse über Itemcharakteristika (z. B. Itemschwierigkeit) haben, die bei der Bestimmung der Testergebnisse mitberücksichtigt werden müssen. CAT birgt hier im Vergleich zum Szenario der mündlichen Prüfung allerdings den entscheidenden Vorteil, dass die subjektiven Eindrücke des Prüfenden bezüglich des Kompetenzniveaus des Prüflings und der Schwierigkeiten der Fragen durch objektive statistische Schätzungen ersetzt werden. Hierfür werden bei CAT Modelle der IRT eingesetzt, auf die im Folgenden näher eingegangen wird.

2.2 Item Response Theory

2.2.1 Modelle der Item Response Theory

Das Kernstück eines adaptiven Tests ist ein psychometrisches Modell, welches die Beziehung des dargelegten Antwortverhaltens (i. S. v. manifesten kategorialen Daten) einer Person zur individuellen latenten Merkmalsausprägung als Wahrscheinlichkeitsfunktion formuliert. Die Wahrscheinlichkeiten bestimmter Antworten werden in IRT-Modellen als Funktion des zu messenden Merkmals (sog. Personenparameter) sowie der psychometrischen Eigenschaften von Items (sog. Itemparameter) modelliert. Bei der Anwendung von IRT-Modellen auf empirische Daten werden die Modellparameter aus den numerisch bewerteten Antwortdaten von Testpersonen geschätzt. Im Zusammenhang der Parameterschätzung spricht man im IRT-Kontext häufig auch von Skalierung, im Sinne der mathematischen Etablierung einer oder mehrerer Skalen, auf der die Testpersonen gemäß ihrer individuellen Merkmalsausprägung lokalisiert werden. IRT-Modelle sind für CAT besonders geeignet, da sie zwei grundlegende Konzepte berücksichtigen. Das erste ist das Konzept der *Invarianz*, als die Annahme, dass die erwarteten individuellen Testergebnisse der Testpersonen unabhängig davon sind, welche Teilmengen an Items sie aus einem kalibrierten Itempool (siehe Kapitel 2.3.1) zur Bearbeitung vorgelegt bekommen. Durch die Schätzung separater Parameter für die Merkmalsausprägung der Testpersonen (Personenparameter) sowie die psychometrischen Eigenschaften der Items (Itemparameter), wird ermöglicht, dass die resultierenden Testergebnisse für interindividuelle Vergleiche genutzt werden können, auch wenn die Testpersonen jeweils andere Items bearbeitet haben. Das zweite Konzept ist das der *Iteminformation*. Die Menge an diagnostischer Information, die die Antwort auf ein bestimmtes Item liefert, variiert je nach Merkmalsausprägung der Testperson und ist für diejenige Person am höchsten, deren Merkmalsausprägung der Schwierigkeit des Items entspricht. Die Itemschwierigkeit und die latente Merkmalsausprägung einer Person sind also auf einer gemeinsamen Skala verortet und können damit in einen direkten Bezug zueinander gebracht werden (Embretson & Reise, 2000). Das Herstellen dieser gemeinsamen Skala ist eine der zentralen Eigenschaften von IRT-Modellen, durch die eine adaptive Itemauswahl ermöglicht wird. Zudem können die Testergebnisse so direkt mit den inhaltlichen Anforderungen von Items verknüpft

werden, wodurch eine kriteriumsorientierte Testwertinterpretation begünstigt wird.

Gängige IRT-Modelle unterscheiden sich in der Anzahl der Parameter, mit denen der funktionale Zusammenhang zwischen Antwortverhalten und Personenmerkmal modelliert wird. Zudem wird unterschieden, ob das Antwortverhalten durch eine oder mehrere latente Merkmale abgebildet wird (sog. mehrdimensionale IRT-Modelle, z. B. Reckase, 2016). Darüber hinaus werden Modelle für dichotome (zwei sich gegenseitig ausschließenden Antwortkategorien, z. B. richtig/falsch) und polytome Daten (mehr als zwei Antwortkategorien, z. B. falsch/teilweise richtig/richtig; z. B. Samejima, 2016) unterschieden. Da mit steigender Modellkomplexität auch die Anforderungen hinsichtlich der Größe der Kalibrierungsstichprobe steigen, bieten sich für Hochschulklausuren aufgrund der üblicherweise vorzufindenden kleinen Stichproben weniger komplexe Modelle aus der Familie der eindimensionalen dichotomen logistischen IRT-Modelle an (z. B. van der Linden, 2016a; Born & Spoden, 2021). Ein allgemeiner Vertreter dieser Modellklasse ist das dreiparametrische logistische Modell (3PL; Birnbaum, 1968). Dieses beschreibt die Wahrscheinlichkeit einer korrekten Antwort $U_{ji} = 1$ einer Person *j* auf das Item *i* als

$$P(U_{ji} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}, \tag{2.1}$$

wobei $\theta_j \in \mathbb{R}$ die individuelle Merkmalsausprägung von Person *j*, und $a_i \in \mathbb{R}^+$ die Itemdiskrimination, $b_i \in \mathbb{R}$ die Itemschwierigkeit und $c_i \in [0,1]$ die untere Asymptote der Lösungswahrscheinlichkeit (auch Pseudorateparameter genannt) von Item *i* darstellen. Aus der in Gleichung 2.1 dargestellten Modellgleichung für das 3PL-Modell lassen sich das zweiparametrische logistische (2PL) und das einparametrische logistische (1PL) Modell ableiten. Das 2PL-Modell leitet sich aus der Annahme ab, dass für alle Items c_i gleich Null ist. Das 1PL-Modell wiederum trifft die zusätzliche Annahme, dass a_i für alle Items gleich einer Konstanten größer Null (üblicherweise $a_i = 1$) ist. Auch wenn das 1PL- und das 2PL-Modell recht strenge Annahmen treffen, sind sie vor allem für Hochschulklausuren praktikable Alternativen zu komplexeren Modellen, da die Schätzung zusätzlicher Itemparameter in kleinen Stichproben problematisch sein kann.

Neben den oben bereits genannten Vorteilen von IRT-Modellen bieten diese zudem auch die Möglichkeit über Equating-Methoden (z. B. Kolen & Brennan, 2014) eine statistische Verbindung zwischen Testzeitpunkten (z. B. Klausurdurchläufe in verschiedenen Studierendenkohorten) herzustellen und somit die einleitend angesprochene Problematik der Kohortenabhängigkeit von Testwertinterpretationen bei Hochschulklausuren aufzulösen (Born & Fink, 2021). Im Folgenden wird daher näher in die Grundlagen des Equatings eingeführt.

2.2.2 Equating

Als Equating bezeichnet man den Prozess der statistischen Adjustierung von Ergebnissen unterschiedlicher Testzusammenstellungen mit dem Ziel der wechselseitigen Austauschbarkeit der

resultierenden Testergebnisse (Kolen & Brennan, 2014). Equating-Methoden können also genutzt werden, um eine statistische Verbindung zwischen den Klausurergebnissen verschiedener Studierendenkohorten herzustellen. Für übliche Klausuren, bei der die Bewertungsskala an der Anzahl der richtig gelösten Aufgaben festgemacht wird, ist diese Verbindung häufig nicht gegeben, sodass sich die Skala zwischen Klausurdurchläufen unterscheidet, wenn nicht dieselbe Menge an Items oder Items mit exakt denselben Itemparametern genutzt werden. Ohne das Herstellen einer statistischen Verbindung sind direkte Vergleiche der Klausurergebnisse dann nicht zulässig (z. B. Born & Fink. 2021). Zur Veranschaulichung kann man sich eine Klausur vorstellen, die in zwei aufeinanderfolgenden Semestern in ein und derselben Lehrveranstaltung durchgeführt wird. Die Klausuren der beiden Semester bestehen jeweils aus 30 zwischen den Semestern variierenden, dichotomen Items. Die Notenvergabe erfolgt anhand der Anzahl richtig gelöster Aufgaben, wobei der jeweils zu erzielende Punktwert zum Erreichen einer bestimmten Notenstufe zwischen den Semestern konstant gehalten wird. Ist die mittlere Schwierigkeit der Items des ersten Semesters nun generell höher als die des zweiten, so ist für das Erreichen der gleichen Notenstufe im ersten Semester ein höheres individuelles Kompetenzniveau notwendig als im zweiten Jahr. Dieses simple Beispiel verdeutlicht, dass ein solches Vorgehen sowohl die Testfairness als auch den Gleichbehandlungsgrundsatz gefährdet. Wie in Kapitel 2.2.1 bereits erwähnt, bietet die IRT hier den großen Vorteil, dass die Ergebnisse von über Semestern variierenden Klausurzusammenstellungen, die auf dem gleichen IRT-skalierten Itempool basieren, bereits auf einer gemeinsamen Metrik verortet sind, auch wenn unterschiedliche Teilmengen an Items bearbeitet wurden. Es muss lediglich sichergestellt werden, dass der Messgegenstand durch die unterschiedlichen Klausurzusammenstellungen in gleicher Weise repräsentiert ist.

Im Rahmen der Konstruktion von Hochschulklausuren wird es allerdings nur in Ausnahmefällen möglich sein, eine große Menge an Items zu konstruieren und an einer großen Stichprobe zu kalibrieren. Realistischer ist es, dass der Itempool über die Zeit wächst, also pro Klausuranwendung dem Itempool neue Items hinzugefügt werden (siehe Kapitel 3.1; siehe auch Frey, Spoden & Born, 2020 für ein Beispiel, bei dem der Itempool über sechs aufeinanderfolgende Klausurdurchläufe stetig ergänzt und die Skala dabei konstant gehalten wurde). Um sicherzustellen, dass die sukzessive neu hinzugefügten Items auf der bereits etablierten Skala verortet werden können, müssen unterschiedliche Klausurzusammenstellungen neben den neuen Items, auch Items enthalten, deren Itemparameterschätzungen bereits auf der etablierten Skala liegen (im Weiteren Linkitems). Dieses Equating-Design wird auch als *Common-Item Equating to a Calibrated Pool* bezeichnet (Kolen & Brennan, 2014). Bei der IRT-Skalierung werden die Itemparameter der Linkitems als bekannt angenommen und auf ihre bisherigen Schätzungen fixiert (sog. fixed parameter calibration; FPC; z. B. Kim, 2006). Die bereits etablierte Skala wird so „aufgespannt“ und die neuen Items auf ihr verortet. Ein solches Vorgehen basiert auf der Annahme, dass die Itemparameter der Linkitems über die Zeit stabil sind und sich abgesehen von geringfügigen

Zufallsschwankungen nicht verändern. Eine signifikante Änderung der Itemparameter über die Zeit wird auch als Itemparameter Drift bezeichnet (IPD; z. B. Goldstein, 1983). Beispielsweise könnten sich die Schwierigkeitsparameter von einem gewissen Anteil an Items deutlich ändern, wenn die Items bekannt geworden sind und die Lösungen zwischen Studierendenkohorten weiterkommuniziert wurden. Da gedriftete Linkitems nicht dafür geeignet sind, die Skala über die Zeit konstant zu halten, sollten diese vor ihrer Fixierung im Rahmen der FPC auf IPD getestet werden.

Im Rahmen von zwei aufeinanderfolgender Hochschulklausuren könnte sich das Equating zusammenfassend nun wie folgt gestalten (für eine ausführliche Darstellung der Equating-Prozedur im Rahmen psychometrisch fundierter Hochschulklausuren siehe Born & Fink, 2021):

Gegeben sei ein Set an Linkitems, das Teil eines bereits kalibrierten Itempools aus der ersten Klausur ist und somit schon über Itemparameterschätzungen verfügt. Dieses Set an Linkitems ist auch Teil der zweiten Klausur. Um zu prüfen, ob die Itemparameter der Linkitems IPD aufweisen, führt man zunächst eine freie (d. h. ohne Fixierung auf die Linkitemparameter) IRT-Skalierung auf Basis der Daten der zweiten Klausur durch. Die so erhaltenen Itemparameter sind aufgrund der oft willkürlichen Festlegung der Metrik der Item- und Personenparameter (z. B. Fixierung des Mittelwerts der Personenparameter auf 0 mit einer Standardabweichung von 1) nicht direkt mit den bereits existierenden Itemparametern aus der ersten Klausur vergleichbar und müssen durch eine Skalentransformation zunächst auf dieselbe Skala gebracht werden. Die Annahme hierbei ist, dass, sofern das jeweilige IRT-Modell (hier 3PL-Modell) Gültigkeit besitzt, sich die Merkmalausprägungen θ auf den beiden Skalen K (erste Klausur) und L (zweite Klausur) um eine lineare Transformation mit den Konstanten A und B unterscheiden:

$$\theta_{Lj} = A\theta_{Kj} + B, \tag{2.2}$$

wobei θ_{Kj} und θ_{Lj} der individuellen Merkmalausprägung von Person j auf den Skalen K und L entsprechen. Die Beziehung der Itemparameter der beiden Skalen stellt sich dann wie folgt dar:

$$b_{Li} = Ab_{Ki} + B, \tag{2.3}$$

$$a_{Li} = \frac{a_{Ki}}{A}, \tag{2.4}$$

und

$$c_{Li} = c_{Ki}. \tag{2.5}$$

Es existieren verschiedene Methoden die Transformationskonstanten A und B auf Basis der geschätzten Itemparameter aus zwei Testanwendungen zu schätzen. Generell unterscheidet man die sogenannten Moment-Methoden (z. B. Mean/Sigma- und Mean/Mean-Transformation; Marco, 1977; Loyd & Hoover, 1980), sowie die Characteristic-Curve-Methoden (z. B. Haebara, 1980;

Stocking & Lord, 1983). Bei den Moment-Methoden werden die Transformationskonstanten über die Mittelwerte und Standardabweichungen der geschätzten Item- beziehungsweise Personenparameter ermittelt, wohingegen bei den Characteristic-Curve-Methoden die Transformationskonstanten über die Differenz der item- (Haebara, 1980) beziehungsweise testcharakteristischen Kurven (Stocking & Lord, 1983) der beiden Skalen K und L bestimmt werden. In Beitrag 2 (siehe Kapitel 3.2) werden die hier genannten Skalentransformationsmethoden im Rahmen der kontinuierlichen Kalibrierung von CAT-Itempools (siehe Kapitel 3.1) in umfassender Weise verglichen.

Nach der Transformation der Itemparameterschätzungen aus Klausur L auf die etablierte Skala K können diese nun auf IPD getestet werden (z. B. mittels Lord's χ^2-Test; Lord, 1980). Items, die signifikanten IPD aufweisen, sollten nicht für die abschließende FPC genutzt werden. In dieser werden nur die Itemparameter der Linkitems, die keinen IPD aufweisen, auf ihre Werte aus der früheren Klausuranwendung fixiert und alle anderen Itemparameter frei geschätzt. Die resultierenden Itemparameter liegen dann auf der bereits etablierten Skala.

Wichtig ist abschließend noch der Hinweis, dass es sich beim Equating ausschließlich um eine statistische Verbindung der Klausurergebnisse, nicht jedoch um eine inhaltliche handelt. Ihr Einsatz ist somit grundsätzlich nur sinnvoll, wenn der Messgegenstand der Klausurzusammenstellungen der gleiche ist und auch im gleichen Maße durch die beiden Klausuren abgebildet ist. Auch die Linkitems sollten daher so gewählt werden, dass sie den Messgegenstand der Gesamtklausur möglichst gut widerspiegeln (Kolen & Brennan, 2014).

2.3 Elementare Bausteine des adaptiven Testens

Neben der Wahl des psychometrischen Modells müssen bei der CAT-Konstruktion Entscheidungen bezüglich sechs elementarer Bausteine von CAT getroffen werden (Frey, 2020). Diese sind (1) der kalibrierte Itempool, (2) die Itemauswahl zu Beginn der Testung, (3) die Schätzmethode der individuellen Merkmalsausprägung, (4) die Itemauswahl während des Tests, (5) der Umgang mit Einschränkungen bei der Itemauswahl, sowie (6) die Kriterien für die Beendigung des Tests. Das Regelsystem, in dem die Bausteine (2) bis (6) spezifiziert sind, wird auch als adaptiver Algorithmus bezeichnet. Im Folgenden werden diese Bausteine jeweils kurz beschrieben und erläutert, wie sie im Hinblick auf typische Hochschulklausuren gestaltet sein sollten.

2.3.1 Kalibrierter Itempool

Der Itempool bezeichnet die Menge an Items, auf die der adaptive Algorithmus während des Tests zugreifen kann. Damit der Algorithmus Items entsprechend der individuellen Merkmalsausprägung automatisiert auswählen kann, müssen die Itemparameter der Items im Itempool bekannt sein. Üblicherweise werden diese im Rahmen von Kalibrierungsstudien ermittelt. Auf Grundlage der Antwortdaten aus der Kalibrierungsstudie werden die Itemparameter auf Basis eines IRT-Modells

geschätzt. Diese Parameter werden in der operationalen CAT-Phase dann als wahre Itemparameter angenommen und für die Personenparameterschätzung (siehe Kapitel 2.3.3) und die adaptive Itemauswahl (siehe Kapitel 2.3.4) genutzt. Das Vorgehen basiert auf der Annahme, dass die Itemparameter invariant für unterschiedliche Populationen von Testpersonen, unterschiedliche Messbedingungen und unterschiedliche Testzusammenstellungen sind (Rupp & Zumbo, 2006). Bezüglich der notwendigen Kalibrierungsstichprobe werden als grobe Daumenregel für das 1PL-Modell ein Minimum von einigen Hundert und für das 2PL-Modell von 500 Antworten je Item angegeben (de Ayala, 2022). Für das 3PL-Modell liegen die Anforderungen an die Stichprobengröße noch deutlich höher. Die Realisierung einer solchen Stichprobengröße für eine Kalibrierungsstudie im Vorlauf der operativen Testanwendung einer Klausur dürfte im Rahmen der üblichen Hochschullehre kaum realisierbar sein. Vielmehr empfiehlt sich eine Kalibrierung der Items im laufenden Lehrbetrieb. Beitrag 1 widmet sich dieser Thematik und schlägt mit der kontinuierlichen Kalibrierungsstrategie (KKS; siehe Kapitel 3.1) eine gangbare Methode für die Kalibrierung von Items für adaptive Hochschulklausuren vor. Die Itemparameterschätzungen werden bei der KKS stetig aktualisiert, nachdem alle Studierenden eines Testzeitpunktes ihre Antworten gegeben haben. Dieses Vorgehen ermöglicht die Kalibrierung neuer Items über mehrere Testanwendungen ohne eine vorgeschaltete Kalibrierungsstudie.

2.3.2 Itemauswahl zu Beginn des Tests

Zu Beginn eines adaptiven Tests liegen noch keine Antworten der Testperson vor. Daher muss die Auswahl des ersten Items anders erfolgen als für den Rest des Tests. Neben einer rein zufälligen Auswahl des ersten Items werden hier auch häufig Items genutzt, die eine mittlere Schwierigkeit aufweisen. Alternativ können auch etwas leichtere Items (sog. Eisbrecheritems) genutzt werden, um den Einstieg in den Test zu erleichtern, und Aspekten wie Prüfungsangst und Nervosität entgegenzuwirken (Frey, 2021). Eine weitere Möglichkeit besteht in der Nutzung von Vorinformationen über die individuelle Merkmalsausprägung (z. B. aus Tests, die ein ähnliches oder hoch korreliertes Merkmal messen). Auf ein solches Vorgehen sollte im Bereich von Hochschulklausuren allerdings verzichtet werden, um den Grundsatz der Chancengleichheit zu wahren.

2.3.3 Schätzung der individuellen Merkmalsausprägung

Bei CAT wird üblicherweise nach der Beantwortung jedes einzelnen Items und nach Abschluss des Tests die individuelle Merkmalsausprägung auf Basis der gegebenen Antworten sowie der jeweiligen Itemparameter geschätzt. Für die Schätzung stehen sowohl Maximum Likelihood Schätzverfahren als auch Bayesianische Verfahren zur Verfügung. Der Unterscheid der beiden Gruppen an Schätzverfahren besteht darin, dass bei Bayesianischen Verfahren Vorannahmen über die Merkmalsverteilung in der Zielpopulation in Form sogenannter Prior-Verteilungen mit in die

Schätzung der individuellen Merkmalsausprägung einbezogen werden. Typische Vertreter der ersten Gruppe sind der Maximum Likelihood Schätzer (MLE; Lord, 1980) sowie der gewichtete Maximum Likelihood Schätzer (engl. weighted maximum likelihood estimator; WLE; Warm, 1989) als Bias korrigierte Variante des Maximum Likelihood Schätzers. Typische Bayesiansiche Vertreter sind der Maximum A Posterioi (MAP; Mislevy, 1986) sowie der Expected A Posteriori (EAP; Bock & Mislevy, 1982) Schätzer (für einen Vergleich verschiedener Personenparameterschätzer bei CAT siehe zum Beispiel Cheng & Liou, 2000; van der Linden & Pashley, 2010). Die Empfehlungen für einen geeigneten Schätzer unterscheiden sich dabei, je nachdem ob es sich um die vorläufige Schätzung während des Tests oder die abschließende Schätzung der Merkmalsausprägung nach Abschluss des Tests handelt. Die Personenparameterschätzung **während** des Tests dient als Basis für die adaptive Itemauswahl. Hier sollten, insbesondere solange das Antwortmuster invariant ist (i. S. v. ausschließlich korrekt oder ausschließlich falsch beantworteten Items), Bayesianische Schätzverfahren genutzt werden, da diese im Gegensatz zum MLE auch für invariante Antwortmuster eine Schätzung liefern. Da das Ziel von Hochschulklausuren generell in der Ableitung rechtsicherer individualdiagnostischer Schlüsse hinsichtlich der in einer Lehrveranstaltung erworbenen Kompetenzen ist, sollten für die **abschließende** Schätzung jedoch auf Bayesianische Verfahren verzichtet werden. Vorannahmen im Sinne von Prior-Verteilungen sind bei adäquater Formulierung auf Gruppenebene korrekt, nicht aber unbedingt auf Individualebene. Um Rechtsunsicherheiten zu vermeiden und die Chancengleichheit der Studierenden zu wahren empfiehlt es sich für adaptive Hochschulklausuren daher den WLE für die abschließende Schätzung zu nutzen (Frey, 2021).

2.3.4 Itemauswahl während der Testung

Im Gegensatz zur Itemauswahl zu Beginn des Tests, kann bei der Itemauswahl während des Tests auf Informationen aus dem bisher gezeigtem Antwortverhalten der Testpersonen zurückgegriffen werden, um das am besten passende Item aus der Menge der noch nicht dargebotenen Items auszuwählen. Die Entscheidung, welches Item gewählt wird, basiert auf einem statistischen Optimalitätskriterium unter Berücksichtigung nicht-statistischer Einschränkungen (siehe Kapitel 2.3.5). Es existieren zahlreiche verschiedene Optimalitätskriterien für die Itemauswahl bei CAT (für eine Übersicht siehe z. B. van der Linden & Glas, 2010), wobei das am häufigsten genutzte Kriterium im Bereich eindimensionaler CATs, welches sich wiederum auch für Hochschulklausuren am besten eignet, das der maximalen Iteminformation ist (Frey, 2021). Bei diesem Optimalitätskriterium wird für jedes noch nicht administrierte Item *i* aus dem Itempool der itemspezifische Wert der Iteminformationsfunktion am Punkt der aktuellen Schätzung für die individuelle Merkmalsausprägung $\hat{\theta}_j$ von Person *j* berechnet und das Item mit dem höchsten Wert ausgewählt. Je nach genutztem IRT-Modell variiert die Berechnung der Iteminformationsfunktion. Für das 3PL-Modell berechnet sich die Iteminformation wie folgt (de Ayala, 2022):

$$I_i(\theta) = {a_i}^2 \left(\frac{P_i(\theta) - c_i}{1 - c_i} \right)^2 \frac{Q_i(\theta)}{P_i(\theta)}, \tag{2.2}$$

wobei $P_i(\theta)$ die Wahrscheinlichkeit Item *i* gegeben der latenten Merkmalsausprägung θ zu lösen und $Q_i(\theta)$ die entsprechende Gegenwahrscheinlichkeit, also die Wahrscheinlichkeit Item *i* gegeben der latenten Merkmalsausprägung θ nicht zu lösen, ausdrücken. Dabei gilt $Q_i(\theta) = 1 - P_i(\theta)$. Für das 2PL-Modell vereinfacht sich Gleichung 2.2 zu

$$I_i(\theta) = {a_i}^2 P_i(\theta) Q_i(\theta), \tag{2.3}$$

und für das 1PL-Modell schließlich zu

$$I_i(\theta) = P_i(\theta) Q_i(\theta). \tag{2.4}$$

Aus Gleichung 2.4 geht hervor, dass für das 1PL-Modell die Iteminformation am höchsten ist, wenn $P_i(\theta) = Q_i(\theta) = 0.5$. Dieser Punkt entspricht dem Punkt, an dem Merkmalsausprägung und Itemschwierigkeit übereinstimmen. Somit wird bei CAT unter Verwendung des 1PL-Modells und dem Auswahlkriterium der maximalen Iteminformation dasjenige Item i ausgewählt, dessen Schwierigkeitsparameter b_i die geringste Differenz zur vorläufigen Merkmalsschätzung $\hat{\theta}_j$ von Person j aufweist. Bei anderen IRT-Modellen, wie dem 2PL oder 3PL-Modell resultiert bei der Itemauswahl nicht zwangsläufig eine solche Passung zwischen Itemschwierigkeit und vorläufiger Personenparameterschätzung, da hier die zusätzlichen Itemparameter mit in die Berechnung der Iteminformationsfunktion eingehen. Vor allem der Diskriminationsparameter hat hier einen entscheidenden Einfluss auf die Itemauswahl, da er in quadrierter Form in die Iteminformation eingeht. Durch den Zusammenhang

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} = \frac{1}{\sqrt{\sum_{i=1}^{k} I_i(\theta)}}, \tag{2.5}$$

gilt jedoch, dass sich der Standardfehler der Merkmalsschätzung unter Nutzung des Auswahlkriteriums der maximalen Iteminformation mit jedem Item maximal verringert, wobei $I(\theta)$ gleich der Testinformation berechnet aus der Summe der Iteminformationen aller k bisher beantworteten Items ist.

2.3.5 Umgang mit Einschränkungen bei der Itemauswahl

Eine Itemauswahl auf Basis der maximalen Iteminformation maximiert die gewonnene diagnostische Information über den Testverlauf. Darüber hinaus sind bei CAT jedoch üblicherweise auch eine Reihe nicht-statistischer Eigenschaften bei der Itemauswahl zu berücksichtigen. Häufig ist daher ein Kompromiss zwischen einem statistischen Optimum und der Erfüllung nicht-statistischer Anforderungen an den Test vonnöten. Methoden zur Berücksichtigung von nicht-

statistischen Testanforderungen werden unter dem Begriff Constraint-Management zusammengefasst. Im Wesentlichen unterscheidet man Methoden zur Kontrolle der Vorgabehäufigkeit von Items und Methoden zur Berücksichtigung inhaltlicher Anforderungen an den Gesamttest.

Die erste Gruppe lässt sich unter dem Begriff der Exposure-Control Methoden subsumieren. Die Verwendung eines rein statistischen Optimalitätskriteriums bei CAT kann dazu führen, dass bestimmte Items sehr vielen Testpersonen und andere Items wiederum so gut wie gar niemanden zur Bearbeitung vorgelegt werden. Eine hohe Vorgabehäufigkeit erhöht das Risiko, dass Items bekannt werden und im Falle von Hochschulklausuren beispielsweise zwischen Kohorten weiterkommuniziert werden. Studentinnen und Studenten hätten so die Möglichkeit zur intensiven Vorbereitung der Lösungen, und die Lösungsquoten würden somit ansteigen (Spoden & Fink, 2021). Damit ein CAT-System über einen langen Zeitraum verwendet werden kann und um die Vergleichbarkeit der Ergebnisse über mehrere Testanwendungen hinweg gewährleisten zu können, gilt es ein solches Sicherheitsrisiko zu vermeiden. Hierfür existieren verschiedene Methoden. Den meisten dieser Methoden ist gemein, dass sie der Itemauswahl eine Zufallskomponente hinzufügen, welche zu einer gleichmäßigeren Ausnutzung des Itempools führt (für einen Vergleich verschiedener Exposure-Control Methoden siehe Leroux et al. 2013).

Die zweite Gruppe an Constraint-Management Methoden dient der Kontrolle von inhaltlichen Anforderungen an den Test auf Individualebene. Mit solchen Content-Management Methoden kann kontrolliert werden, wie hoch der Anteil an Items pro Inhaltsbereich, pro Stimulusformat (z. B. Text, Bild, Video), pro Antwortformat (z. B. offen, geschlossen) und ähnliches am Gesamttest ausfallen soll. Der Shadow-Testing Ansatz von van der Linden und Reese (1998) ist hier besonders gut geeignet. Dieser ermöglicht die gleichzeitige Berücksichtigung einer hohen Anzahl nicht-statistischer Einschränkungen und liefert sehr gute Ergebnisse bezüglich der Constraint-Verletzungen (für ein Review zum Shadow-Testing Ansatz siehe van der Linden, 2021). Darüber hinaus bietet der Shadow-Testing Ansatz die Möglichkeit, Exposure-Control Methoden und Content-Management Methoden in einen gemeinsamen methodischen Rahmen zu integrieren (van der Linden & Choi, 2019; van der Linden & Veldkamp, 2004, 2007). Auch wenn methodisch anspruchsvoller in der Implementation als beispielsweise heuristische Content-Management Methoden (für einen Vergleich verschiedener heuristischer Verfahren siehe z. B. Born & Frey, 2017), ist es daher auch für Hochschulklausuren ratsam, den Shadow-Testing Ansatz als Constraint-Management Methode zu nutzen.

2.3.6 Kriterien für Beendigung des Tests

Die adaptive Itemauswahl bei CAT wird so lange fortgesetzt, bis ein oder mehrere vorab definierte Kriterien für die Beendigung des Tests erreicht sind. Häufig verwendete Kriterien für die Beendigung des Tests sind a) das Erreichen einer festgesetzten maximalen Anzahl an zu bearbeitenden Items,

b) das Erreichen einer Maximalen Testzeit und c) das Überschreiten einer festgelegten Präzision für die Schätzung der Merkmalsausprägung (z. B. ein ausreichend kleiner Standardfehler der Merkmalsschätzung), sowie etwaige Kombinationen dieser Kriterien (für einen Vergleich verschiedener Abbruchkriterien siehe Babcock & Weiß, 2012). Letzteres resultiert allerdings in adaptiven Tests mit individuell unterschiedlichen Testlängen. Kleine Itempools oder Itempools mit wenigen oder unzureichend vielen Items an den Rändern der Merkmalsverteilung können dazu führen, dass für Personen in den Randbereichen das Kriterium einer definiert hohen Messpräzision nicht oder erst nach Beantwortung sehr vieler Items erreicht wird. Aus diesem Grund und da Klausuren im Hochschulbereich typischerweise als Gruppentestungen durchgeführt werden, empfiehlt es sich auch aus logistischen Gründen hier als Abbruchkriterium die maximale Testzeit zu nutzen. Um zu vermeiden, dass schnell arbeitende Studierende sehr viele Items bearbeiten müssen, empfiehlt sich zudem die Kombination mit einer maximalen Itemanzahl (Frey, 2021).

2.4 Vorteile von CAT für Hochschulklausuren

2.4.1 Messeffizienz

Aus statistischer Sicht besteht der Hauptvorteil von CAT gegenüber nicht-adaptiven Tests in der Möglichkeit, die Messeffizienz der Testverfahren erheblich zu steigern (Segall, 2005). Dieser Effizienzgewinn kann einerseits zur Erhöhung der Messgenauigkeit genutzt werden, wenn die Anzahl der Items für alle Testteilnehmer konstant gehalten wird, oder zur Reduzierung der Testlänge. Im Vergleich zu nicht-adaptiven Tests kann die Anzahl der Items beim Einsatz von CAT typischerweise um etwa die Hälfte reduziert werden (bei vergleichbarer Messgenauigkeit; Segall, 2005). Darüber hinaus bietet CAT die Möglichkeit, das Problem zu überwinden, dass konventionelle Tests typischerweise im mittleren Bereich der Merkmalsverteilung viel genauer messen als bei extremer Merkmalsausprägung (z. B. de Ayala, 2022). Dies wird durch die Angleichung der Standardfehler der Merkmalsschätzungen über die Merkmalsverteilung erzielt. So kann erreicht werden, dass die aus der Klausurbearbeitung abgeleiteten Bewertungen in Form von beispielsweise Noten, über die komplette Notenskala die gleiche Genauigkeit aufweisen (z. B. Weiss & Kingsbury, 1984). Eine notwendige Voraussetzung hierfür ist ein hinreichend großer Itempool, der für alle Merkmalsausprägungen genügend Items beinhaltet. Am optimalsten wäre ein Itempool mit gleichverteilter, hoher Testinformation über alle in der Stichprobe der Testpersonen vorhandenen Merkmalsausprägungen (Segall, 2005).

2.4.2 Validität

Validität als Gütekriterium bezieht sich auf die Gültigkeit von Interpretationen von Testwerten für die beabsichtigten Verwendungen eines Tests (Hartig et al., 2020). Als Evidenz für die Validität werden Informationen aus verschiedenen Quellen genutzt. Unter diesen sind üblicherweise Untersuchungen, die überprüfen, inwieweit die ermittelten Testergebnisse inhaltlich erwartete

Zusammenhänge zu auf andere Weise erhobenen Testwerten für das gleiche beziehungsweise theoretisch eng verwandte Merkmale (sog. konvergente Evidenz), sowie keinen Zusammenhang zu theoretisch abgrenzbaren Merkmalen aufweisen (sog. diskriminante Evidenz; Frey, 2020). Oben genannte Möglichkeit den Messeffizienzvorteil von CAT für die Erhöhung der Messpräzision zu nutzen, was wiederum gleichbedeutend mit der Erhöhung des Anteils systematischer, auf die Merkmalsausprägung zurückzuführender Varianz der Personenparameterschätzung in Relation zur Fehlervarianz ist, lässt höhere statistische Zusammenhänge mit konvergenten Variablen erwarten (Frey, 2006). Mit Bezug auf die diskriminante Evidenz ist die Abschätzung der Effekte von CAT allerdings nicht so einfach. Empirische Studien kamen hier zu gegenteiligen Ergebnissen (Moosbrugger & Goldhammer, 2007; Ortner & Caspers, 2011). Etwaig verzerrende Zusammenhänge mit diskriminanten Variablen wie Testangst oder Intelligenz können allerdings durch eine transparente Erläuterung der Funktionsweise des adaptiven Tests in der Testinstruktion vermieden werden (Ortner & Caspers, 2011). Eine Diskussion dazu, wie der Einfluss konstruktirrelevanter Faktoren auf die Testergebnisse durch computerbasiertes Testen sowie CAT kontrolliert werden kann, findet sich bei Wise (2019). Sieht man CAT als ein System, dessen erklärtes Ziel es ist, sich dem Verhalten von Testteilnehmern anzupassen und somit die Testeffizienz und die Validität der Testwertinterpretationen zu erhöhen, so kann darüber hinaus argumentiert werden, dass ein CAT auch derart gestaltet werden könnte, dass er sich nicht nur dem Kompetenzniveau der Testpersonen anpasst, sondern auch auf anderes Testteilnehmerverhalten adaptiv reagiert. So könnte beispielsweise auf verzerrende Effekte durch unmotivierte Testbearbeitung (z. B. detektiert durch extrem kurze Antwortzeiten; z. B. Wise & Gao, 2017) reagiert werden, indem der CAT die Antwortzeiten überwacht, und sollte er unmotiviertes Antwortverhalten detektieren, darauf adaptiv reagiert (z. B. durch das Einblenden einer Nachricht im Testsystem; für ein Beispiel eines solchen Tests und dessen positive Effekte auf Testanstrengung, Testleistung und Validität siehe Kong et al., 2006; Wise et al., 2006). Intelligente CAT-Systeme, die sich auch anderen Verhaltensmarkern als den reinen Itemantworten anpassen, haben somit das Potential die Validität der aus den Testergebnissen abgeleiteten Interpretationen zu steigern (Wise, 2020).

2.4.3 Testsicherheit

Neben den psychometrischen Vorteilen kann die erhöhte Sicherheit von adaptiven Klausuren als weiterer Vorteil angeführt werden. Das Prinzip des adaptiven Testens beinhaltet die individualisierte Testzusammenstellung anhand des gezeigten Antwortverhaltens der jeweiligen Testperson. Individualisierte Klausuren können als relativ sicher angesehen werden, da eine individualisierte Klausurzusammenstellung die Möglichkeit des Abschreibens von Aufgabenlösungen beziehungsweise ein Erschleichen von Vorteilen aufgrund vorheriger Kenntnis

von Klausuraufgaben (z. B. aus Gedächtnisprotokollen von Studierenden höherer Semester) erheblich erschwert (Spoden & Fink, 2020).

2.4.4 Effekte auf Motivation

Insbesondere der Umstand, dass bei CAT alle Testpersonen nur auf ihr individuelles Kompetenzniveau abgestimmte Items präsentiert bekommen, lässt vermuten, dass die Art des Testalgorithmus einen Einfluss auf die Motivation zur Testbearbeitung und somit wiederum auf die Testleistung hat. Die Befundlage hierzu ist nicht eindeutig. Vor allem frühere Arbeiten lassen auf eine motivationssteigernde Wirkung von CAT schließen (Betz, 1975; Betz & Weiss 1976a; 1976b, Pine et al., 1979). Dies wurde üblicherweise damit erklärt, dass bei der adaptiven Itemauswahl vermieden werden würde, dass Testpersonen viel zu leichte oder viel zu schwere Items bearbeiten müssen. Allerdings existiert wenig empirische Evidenz für diese Annahmen (Wise, 2014). Neuere Arbeiten weisen dagegen auf eine motivationsmindernde Wirkung von CAT hin. So wird argumentiert, dass CAT im Allgemeinen als zu schwierig wahrgenommen werden könnte, da Testteilnehmer an traditionelle Leistungstests mit fester Itemreihenfolge und höherer Lösungswahrscheinlichkeit gewöhnt sind. Diese Abweichung vom gewohnten Testerleben kann wiederrum einen negativen Einfluss auf die Motivation haben (z. B. Bergstrom et al., 1992; Eggen, 2004; Eggen & Verschoor, 2006; Ponsoda et al., 1999). Gerade für leistungsfähige Personen könnte die ungewohnt niedrige Lösungswahrscheinlichkeit demotivierend wirken, wobei angenommen wird, dass diese Effekte stärker sind, als die oben erwähnten motivationssteigernden Effekte auf leistungsschwache Testpersonen (Frey, 2020). Die Position einer motivationsmindernden Wirkung von CAT wird von verschiedenen experimentellen Untersuchungen gestützt (Frey et al., 2009; Ortner et al., 2014; Tonidandel et al., 2002). Eine Integration beider Positionen lässt sich bei Asseburg (2011) finden. Sie stützte sich in einer umfassenden experimentellen Untersuchung auf das Erwartungs-Wert-Modell der Motivation zur Testbearbeitung. Ihre Ergebnisse verdeutlichen, dass sich der Einsatz von CAT in Abhängigkeit bestimmter Situations-, Test- und Personenmerkmale unterschiedlich auf die Motivation zur Testbearbeitung auswirkt. Diese differentiellen Zusammenhänge könnten eine Erklärung für die oben geschilderten inkonsistenten Ergebnisse bisheriger empirischer Studien zum Effekt von CAT auf die Motivation liefern. Zahlreiche Wechselwirkungen verschiedener Persönlichkeits- und Testmerkmale mit der Motivation zur Testbearbeitung und daraus resultierende, zwischen Personengruppen variierende, motivationale Auswirkungen von CAT können allerdings durch eine transparente Erläuterung der Funktionsweise eines adaptiven Tests in der Testinstruktion vermieden werden (Asseburg, 2011). Diese Ergebnisse sowie weitere Forschungsarbeiten (z. B. Ortner & Caspers, 2011) lassen in Bezug auf die Konstruktion adaptiver Hochschulklausuren schließen, dass für eine faire Klausur sowie diskriminant valide Testwertinterpretationen den Studierenden die Funktionsweise eines

adaptiven Tests in der Klausurinstruktion erläutert werden sollte. Zudem sind auch moderne CAT-Systeme, die sich unmotiviertem Testverhalten anpassen, denkbar (siehe Kapitel 2.4.2).

2.4.5 Vorteile von E-Klausuren

Neben den genannten CAT-spezifischen Vorteilen bieten computerisierte adaptive Hochschulklausuren auch Vorteile, die sich aus der Nutzung von E-Klausuren im Allgemeinen ableiten lassen. Unter diese fallen unter anderem die Zeit- und Ressourcenersparnis vor allem durch die Möglichkeit zur automatisierten Testzusammenstellung, sowie Auswertung und Analyse der Klausurdaten, die Möglichkeit zur direkten Rückmeldung der Klausurergebnisse nach Beendigung des Tests, die Möglichkeit zur Integration innovativer und authentischerer Items (z. B. durch die Nutzung von Multimediadateien oder interaktive Items), die Möglichkeit zum automatischen Aufzeichnen von Log-Daten für weitergehende Analysen und viele weitere (z. B. Boevé et al., 2015; Nikou & Economides, 2018; Rolim & Isaias, 2019; Spoden & Frey, 2021, St-Onge et al., 2021). Darüber hinaus kann sich der Einsatz von computerbasierten im Vergleich zu herkömmlichen papierbasierten Tests positiv auf die Motivation zur Testbearbeitung die Selbstwirksamkeit, die Testwahrnehmung und sogar die Testleistung auswirken (Chua & Don, 2013; Gu et al., 2020; Nardi & Ranieri, 2018; Nikou & Economides, 2016; Rolim & Isaias, 2019). Zudem lassen sich durch die Nutzung des Computers auch realitätsnähere Klausuren gestalten, in denen die Studierenden mit Hilfe fachspezifischer Standardsoftware Aufgaben bearbeiten (z. B. die Nutzung von Statistiksoftware zur Bearbeitung von Items in einer Statistikklausur).

3 Herausforderungen bei der Implementation von kriteriumsorientierten adaptiven Hochschulklausuren – Darstellung der Einzelbeiträge

Auch wenn CAT ein großes Potential besitzt Hochschulklausuren individualisierter, fairer und messgenauer zu gestalten, ist die reine Bewusstwerdung der Vorteile kein Garant für dessen Nutzung. Vielmehr müssen, wie einleitend bereits erläutert, für eine zielführende und flächendeckende Implementation von kriteriumsorientierten, adaptiven Hochschulklausuren zunächst verschiedene Herausforderungen auf psychometrischer, personaler und technischer Ebene gemeistert werden. Die Überwindung dieser ist übergeordnetes Ziel dieser Arbeit. Die hier vorliegende kumulative Dissertation umfasst vier wissenschaftliche Artikel, welche im Folgenden zusammenfassend dargestellt werden. Die vollständigen Artikel finden sich im Anhang (Anhang A – D).

3.1 Beitrag 1: A continuous calibration strategy for computerized adaptive testing

Zitation: Fink, A., Born, S., Spoden, C. & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling,* 60, 327–346. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_3-2018_327-346.pdf

3.1.1 Einleitung

Ein essentieller Bestandteil von CAT ist ein hinreichend großer Itempool mit gut geschätzten Itemparametern, der üblicherweise in separaten Kalibrierungsstudien vor der eigentlichen Testanwendung kalibriert wird (siehe Kapitel 2.3.1). In vielen potentiellen Anwendungsbereichen von CAT (z. B. Hochschulklausuren) ist die massierte Konstruktion des gesamten Itempools sowie die Durchführung einer separaten Kalibrierungsstudie aufgrund mangelnder Ressourcen für die Testentwicklung nicht realisierbar. Eine praxistaugliche Lösung dieser Problematik liefert dabei die Methode der Online-Kalibrierung (z. B. Stocking, 1988). Ursprünglich wurde diese Methode entwickelt, um neue Items während der Administration eines CATs zu kalibrieren. Die meisten der existierenden Online-Kalibrierungsmethoden basieren auf der Annahme, dass bereits ein betriebsbereiter CAT oder zumindest eine gewisse Menge bereits kalibrierter Items vorhanden ist (z. B. Verschoor et al., 2019). Die Online-Kalibrierung wird dann genutzt, um neue Items dem bereits existierenden Itempool hinzuzufügen. Es finden sich allerdings bisher nur wenige Studien, die sich mit Online-Kalibrierungsmethoden beschäftigen, um einen Itempool zu kalibrieren, über dessen Parameter zu Beginn des Testeinsatzes keine Informationen vorliegen. Makransky und Glas (2010)

entwickelten eine Methode, bei der die Itemparameter nach jeder Administration der Items unter Verwendung aller bisherigen Antworten neu geschätzt werden. Somit wird die Präzision der geschätzten Itemparameter und darauf aufbauend auch die Personenparameterschätzungen kontinuierlich über die Zeit verbessert.

Aus zwei Gründen kann diese Strategie allerdings problematisch für Anwendungsbereiche wie Hochschulklausuren sein. Erstens führt das Neuschätzen der Itemparameter nach jeder Itemvorgabe vor allem im high-stakes Bereich zu rechtlichen Problemen, da die Schätzungen der Personenparameter für jede Testperson auf unterschiedlichen Itemparametern basiert und sie somit nicht auf der gleichen, etablierten Skala verortet sind. Zweitens ist es in einigen Anwendungsbereichen notwendig, dass die Skalen über verschiedene Testzyklen hinweg statistisch verbunden werden. Als Testzyklus wird in diesem Zusammenhang die komplette Testprozedur verstanden, welche Schritte wie Testzusammenstellung, Testadministration sowie die Analyse der Testdaten umfasst. Durch die statistische Verbindung kann die Vergleichbarkeit die Ergebnisse über die Testzyklen hinweg gewährleistet werden (siehe Kapitel 2.2.2).

Das Anliegen des ersten Einzelbeitrags ist es daher, eine neue kontinuierliche Kalibrierungsstrategie (KKS) für die Verwendung von CAT bei wiederkehrenden Testanwendungen zu präsentieren und auf Grundlage einer Simulationsstudie ihre Funktionsweise genauer zu untersuchen. Die Hauptfunktionen dieser Kalibrierungsstrategie sind (a) die Nutzung von Itemantworten mehrerer Testzyklen zur Kalibrierung mit einem Modell der IRT; (b) die Beibehaltung der Berichtsmetrik über Testzyklen hinweg durch die Nutzung von Equating-Methoden; (c) die Steigerung der Adaptivität des Tests sowie der Präzision der Personen- und Itemparameterschätzungen über Testzyklen bei gleichzeitiger Kontrolle von (d) IPD. Ein mit der KKS konstruierter Test, beginnt in einer nicht-adaptiven Weise und wird über die Testzyklen hinweg immer adaptiver. Eine vorgeschaltete separate Kalibrierungsstudie ist nicht erforderlich.

3.1.2 Kontinuierliche Kalibrierungsstrategie

In der KKS werden drei Arten von Itemclustern unterschieden die jeweils Items mit bestimmten Eigenschaften enthalten und die einem bestimmten Ziel dienen: Das *Kalibrierungsscluster*, das *adaptive Cluster* und das *Linkingcluster.* Das Kalibrierungscluster enthält ausschließlich unkalibrierte Items, die dazu dienen den Itempool in der ersten Testanwendung zu etablieren und in folgenden Testanwendungen durch das hinzufügen neuer Items zu vergrößern. Das adaptive Cluster enthält Items, die unter Verwendung der Itemparameterschätzungen aus dem vorherigen Testzyklus adaptiv gewählt werden, um so die Messgenauigkeit des Tests zu erhöhen. Das Linkingcluster dagegen enthält Items, die zum Linking zweier aufeinanderfolgender Testzyklen im Sinne eines Common Item Nonequivalent Group Designs (siehe Kapitel 2.2.2) genutzt werden, um die Berichtsmetrik über Testanwendungen stabil zu halten. Die Items im Kalibrierungs- und im Linkingcluster werden allen Testpersonen eines Testzyklus präsentiert. Ein Test kann je nach

gewünschtem Grad an Adaptivität, Menge an Linkitems oder Menge an neuen Items mehr oder weniger der entsprechenden Cluster enthalten. Wird die KKS in Situationen genutzt, in denen keine Itemparameter zu Beginn der Prozedur vorliegen, dann umfasst der erste Testzyklus ausschließlich Kalibrierungscluster. Ab dem zweiten Testzyklus setzen sich die Tests aus allen drei Clusterarten zusammen. Der Algorithmus der KKS setzt sich dann aus folgenden sieben Schritten zusammen (siehe Abbildung 1 im Vollbeitrag in Anhang A für ein Flussdiagram der KKS):

1. Auswahl der Linkitems für das Linkingcluster.
2. Testspezifikation und Testadministration.
3. Freie Skalierung aller Items mit Daten des aktuellen Testzyklus.
4. Skalentransformation der Linkitems.
5. Inferenzstatistischer Test auf IPD. Linkitems, die signifikanten IPD zeigen, werden in Schritt 6 aus der FPC ausgeschlossen und frei geschätzt.
6. FPC auf Basis der Daten aller bisherigen Testzyklen im Sinne eines concurrent Scalings. Fixierung der Parameter der nicht-gedrifteten Linkitems auf ihre Schätzwerte aus dem vorherigen Testzyklus.
7. Personenparameterschätzung unter Verwendung der aus Schritt 6 resultierenden Itemparameterschätzungen.

3.1.3 Fragestellungen

Neben der Präsentation der KKS ist ein Hauptziel der Studie die Leistungsfähigkeit der KKS hinsichtlich der Qualität der Personenparameterschätzungen zu untersuchen, sowie praktische Empfehlungen hinsichtlich der Konfiguration des Algorithmus zu geben. Folgende Fragestellungen wurden mit der Studie untersucht:

1. Wie wirken sich unterschiedliche Stichprobengrößen pro Testzyklus während der KKS auf die Präzision der Personenparameterschätzungen nach unterschiedlicher Anzahl an Testzyklen aus?
2. Wie wirkt sich die Geschwindigkeit bei der Kalibrierung neuer Items während der KKS auf die Präzision der Personenparameterschätzungen nach unterschiedlicher Anzahl an Testzyklen aus?
3. Inwieweit beeinflusst das gewählte IRT-Modell bei der KKS die Ergebnisse für Forschungsfragen 1 und 2?

3.1.4 Methode

Zur Beantwortung der oben genannten Forschungsfragen wurde eine Monte-Carlo-Simulation auf der Basis eines dreifaktoriellen Designs mit den Faktoren *Stichprobengröße pro Testzyklus* (N = 50, N = 100, N = 300), *Kalibrierungsgeschwindigkeit* (t = 3, t = 5, t = 9) und *IRT-Modell* (1PL, 2PL)

durchgeführt. Die Faktorstufen des Faktors Kalibrierungsgeschwindigkeit repräsentieren die notwendigen Testzyklen *t,* um die angestrebte Itempoolgröße von 130 kalibrierten Items zu erreichen. Sie wird durch die Anzahl an Items im Kalibrierungscluster determiniert. Die Testlänge pro Testzyklus betrug 50 Items und die Menge an Linkitems ab dem zweiten Testzyklus wurde auf 10 fixiert. Das bedeutet, für die Faktorstufe *t* = 9 umfasste jeder Test ab dem zweiten Testzyklus 30 Items im adaptiven Cluster und 10 Items im Kalibrierungscluster. Da sich der Itempool in jedem Testzyklus um 10 Items vergrößert, wurde der gesamte Itempool in dieser Bedingung nach neun Testzyklen kalibriert. Eine Übersicht der Testzusammenstellungen getrennt nach den Faktorstufen des Faktors Kalibrierungsgeschwindigkeit ist in Tabelle 1 dargestellt. Für jede Bedingung lief der KKS-Algorithmus für neun Testzyklen. Das vollständig gekreuzte Design resultierte in 3 x 3 x 2 = 18 Bedingungen. Als Evaluationskriterien für die Messpräzision dienten der globale Mean Squared Error (*MSE*; mittlere quadratische Abweichung der geschätzten von den wahren Personenparametern) nach jedem Testzyklus, sowie der auf den wahren Personenparameter bedingte *MSE* nach jedem Testzyklus. Darüber hinaus wurden eine obere (Personenparameterschätzung auf Basis der wahren Itemparameter) und eine untere Basisbedingung (Personenparameterschätzung auf Basis von Itemparametern, die nach ihrer ersten Schätzung fixiert und somit nicht kontinuierlich aktualisiert wurden) simuliert, mit der die Ergebnisse verglichen wurden. Für jede der Bedingungen wurden 200 Replikationen im Hinblick auf

Tabelle 3.1

Itemanzahl pro Cluster und Testzyklus in der kontinuierlichen Kalibrierungsstrategie getrennt nach Kalibrierungsgeschwindigkeit

Kalibrierungs-geschwindigkeit	Cluster	Testzyklus								
		1	2	3	4	5	6	7	8	9
	Adaptiv	0	0	0	40	40	40	40	40	40
t = 3	Kalibrierung	50	40	40	0	0	0	0	0	0
	Linking	0	10	10	10	10	10	10	10	10
	Adaptiv	0	20	20	20	20	40	40	40	40
t = 5	Kalibrierung	50	20	20	20	20	0	0	0	0
	Linking	0	10	10	10	10	10	10	10	10
	Adaptiv	0	30	30	30	30	30	30	30	30
t = 9	Kalibrierung	50	10	10	10	10	10	10	10	10
	Linking	0	10	10	10	10	10	10	10	10

Anmerkung. Testlänge pro Testzyklus = 50; Kalibrierungsgeschwindigkeit = notwendige Anzahl an Testzyklen *t*, um die angestrebte Itempoolgröße von 130 kalibrierten Items zu erreichen.

die Evaluationskriterien analysiert. Für jede Replikation wurden Schwierigkeitsparameter b_i aus einer trunkierten Normalverteilung mit $b_i \sim N(0, 1.5)$, $b_i \in (-4.5, 4.5)$ und Diskriminationsparameter a_i aus einer lognormal-Verteilung mit $a_i \sim logN(0, 0.25)$ gezogen. Personenparameter wurden jeweils aus einer Standardnormalverteilung $\theta \sim N(0,1)$ gezogen. Die Simulation wurde in R (R Core Team, 2021) unter Nutzung der Pakete mirt (Chalmers, 2012) und mirtCAT (Chalmers, 2016) durchgeführt.

3.1.5 Ergebnisse

Da sich die Ergebnisse bezüglich der Fragestellungen 1 und 2 kaum zwischen dem 1PL- und dem 2PL-Modell unterscheiden, wird zur Beantwortung der ersten beiden Fragenstellungen zunächst nur auf die Ergebnisse des 2PL-Modells eingegangen (für die Ergebnisse zum 1PL siehe den vollständigen Beitrag 1 im Anhang A). Mit Bezug auf Fragestellung 1 ist erwartungskonform festzustellen, dass die mittlere Präzision der Personenparameterschätzungen mit zunehmender Stichprobengröße steigt, jedoch mit mittleren *MSE*-Werten (gemittelt über alle Testzyklen) von 0.144 (t = 3), 0.117 (t = 5) und 0.117 (t = 9) bereits bei N = 50 als akzeptabel anzusehen ist. Im Verlauf der Testzyklen konvergiert der MSE zur unteren Basisbedingung und entfernt sich von der oberen Basisbedingung. Der Unterschied zwischen KKS und den beiden Basisbedingungen verschwindet jedoch mit steigender Stichprobengröße. Bezüglich des auf das individuelle Merkmalsniveau bedingten *MSE*s ergab sich für jede Bedingung und in allen Testzyklen im Kalibrierungsprozess die höchste Messpräzision im mittleren Merkmalsbereich. Über den Verlauf der Testzyklen gleicht sich die Messpräzision zwischen Testteilnehmern mit extremer Merkmalsausprägung und Testteilnehmern im mittleren Bereich jedoch an.

Bezüglich Fragestellung 2 erwies sich eine langsame Kalibrierung als vorteilhaft, da die über alle Stichprobengrößen gemittelte Präzision bei der neunten Erhebung geringfügig höher (*MSE* = 0.086; t = 9) als bei mittelschneller (0.096, t = 5) und schneller Kalibrierung (0.099, t = 3) ausfällt. Zudem kommt es bei einer langsamen Kalibrierung zu einem gleichmäßigeren Anstieg der Messpräzision über den Verlauf der Testzyklen.

Mit Bezug auf Fragestellung 3 kann festgestellt werden, dass sich die Ergebnisse bezüglich Fragestellung 1 und 2 kaum zwischen dem 1PL- und dem 2PL-Modell unterscheiden. Lediglich bei sehr kleinen Stichproben und einer schnellen Kalibrierungsgeschwindigkeit (t = 3) ergab sich in den ersten Testzyklen eine höhere Messpräzision für das 1PL-Modell im Vergleich zum 2PL-Modell. Hier scheint die hohe Unsicherheit in den Diskriminationsparametern zu Beginn des Prozesses, den Vorteil der höheren Testinformation des 2PL-Modells zu negieren, sodass in den ersten Testzyklen die Messpräzision sogar etwas abnimmt. Mit Einsatz des adaptiven Testteils (ab dem vierten Testzyklus in der Bedingung t = 3) steigt die Messpräzision aber auch hier deutlich.

3.1.6 Diskussion

Die KKS erzielt selbst für sehr kleine Stichproben von *N* = 50 und unter Verwendung des 2PL-Modells sehr gute Ergebnisse. Die Testungen starten in der ersten Erhebungsrunde nicht-adaptiv und werden über die Testanwendungen zunehmend adaptiver und präziser, wobei Parameterschätzungen über die Zeit optimiert, defizitäre Items identifiziert und die Ergebnisse der Testanwendungen statistisch verbunden werden. Die KKS ist somit eine vielversprechende Methode, wenn das gleiche Konstrukt über mehrere Erhebungen hinweg gemessen werden soll. Für Anwender vermutlich oftmals wünschenswert ist, dass die beste Performanz erzielt wird, wenn je Erhebungsrunde nur wenige Items ergänzt werden. Die KKS erschließt somit die Anwendbarkeit von CAT auch für Bereiche, in denen die Konstruktion einer großen Menge an Items und/oder die Durchführung einer separaten Kalibrierungsstudie mit einer großen Stichprobe vor der eigentlichen Testanwendung aufgrund mangelnder Ressourcen schlichtweg nicht möglich sind.

3.2 Beitrag 2: Evaluating different equating setups in the continuous item pool calibration for computerized adaptive testing

Zitation: Born, S., Fink, A., Spoden, C. & Frey, A. (2019). Evaluating different equating setups in the continuous item pool calibration for computerized adaptive testing. *Frontiers in Psychology*, *10*, 1277. https://doi.org/10.3389/fpsyg.2019.01277

3.2.1 Einleitung

Die in Beitrag 1 ausführlich beschriebene KKS ermöglicht einen schrittweisen Aufbau des Itempools über mehrere Testzyklen ohne separate Kalibrierungsstudie. Da die Itemparameterschätzungen bereits vorhandener und neuer Items über Testanwendungen hinweg kontinuierlich aktualisiert werden, dabei aber anzunehmen ist, dass sich die Merkmalsverteilungen der Testpersonen der verschiedenen Testzyklen unterscheiden können, bildet das der KKS inhärente Equating einen kritischen Einflussfaktor auf die Vergleichbarkeit der Testergebnisse über Testzyklen hinweg. Wie bereits erwähnt, basiert das in der KKS implementierte Equating auf einem Common Item Nonequivalent Group Design mit vier Schritten: (1) Auswahl der Linkitems; (2) Skalentransformation; (3) Detektion von IPD und schließlich (4) Fixierte Skalierung. Die Ergebnisse von Beitrag 1 liefern zwar bereits vielversprechende Ergebnisse bezüglich der Performanz der KKS, lassen jedoch zwei zentrale Punkte außer Acht, die in dieser Studie detailliert untersucht werden. Als erstes ist anzumerken, dass in Beitrag 1 idealisierte Bedingungen vorausgesetzt wurden, indem die Merkmalsverteilungen der Testpersonen sich nicht zwischen den Testzyklen unterschieden. Darüber hinaus wurden die Auswirkungen verschiedener Konfigurationen der Equating-Prozedur (z. B. Schwierigkeitsverteilung der Linkitems, Methode der Skalentransformation) auf die Performanz der KKS nicht untersucht.

Das Ziel von Beitrag 2 ist es daher, die Leistungsfähigkeit der KKS in Verbindung mit verschiedenen Konfigurationen der Equating-Prozedur unter realistischeren Bedingungen (d. h. variierenden mittleren Kompetenzniveaus und Varianzen der Kompetenzverteilungen zwischen Testzyklen) zu untersuchen. Basierend auf den Ergebnissen sollen Empfehlungen für die Konfiguration der KKS abgeleitet werden.

3.2.2 Fragestellungen

1. Welchen Einfluss hat die Schwierigkeitsverteilung der Linkitems in der KKS auf die Messpräzision der Itemparameterschätzungen?
2. Welchen Einfluss hat die Schwierigkeitsverteilung der Linkitems in der KKS auf die Qualität des Equatings?
3. Welchen Einfluss hat die Methode der Skalentransformation in der KKS auf die Qualität des Equatings?

3.2.3 Methode

Zur Beantwortung der Forschungsfragen wurde eine Monte-Carlo-Simulation basierend auf einem dreifraktionellen Design durchgeführt. Mit dem ersten Faktor *Schwierigkeitsverteilung der Linkitems* (normal, gleichverteilt, bimodal), wurde die Verteilung der Leichtigkeitsparameter d_i der Linkitems variiert. Es sei darauf hingewiesen, dass in Beitrag 2 eine etwas andere Parametrisierung der Itemparameter genutzt wurde als in Beitrag 1. So wurden Leichtigkeitsparameter d_i anstelle von Schwierigkeitsparametern b_i verwendet. Aufgrund der geläufigeren Bezeichnung wird mit Bezug auf diesen Faktor dennoch von der Schwierigkeitsverteilung gesprochen. Als zweiter Faktor wurde die im Equatingprozess genutzte Methode der *Skalentransformation* (Mean/Mean, Mean/Sigma, Haebara, Stocking-Lord) variiert. Schließlich wurde mit dem dritten Faktor die *Stichprobengröße* (*N* = 50, *N* = 100, *N* = 300) pro Testzyklus während der kontinuierlichen Kalibrierung variiert. Das vollständig gekreuzte Design resultierte in 3 x 4 x 3 = 36 Bedingungen. Die simulierte Testlänge pro Testzyklus betrug 60 Items (Itemanzahl Linkingcluster = 15, Itemanzahl Kalibrierungscluster = 20, Itemanzahl adaptives Cluster = 25). In jeder Bedingung lief der KKS-Algorithmus für zehn Testzyklen, was zu einem finalen Itempool von 240 Items führte. Für jede der Bedingungen wurden 200 Replikationen analysiert. Als IRT-Modell wurde das 2PL-Modell zugrunde gelegt. Leichtigkeitsparameter d_i wurden pro Replikation zufällig aus einer trunkierten Normalverteilung mit $d_i \sim N(0, 1.5)$, $d_i \in (-2.5, 2.5)$ und Diskriminationsparameter a_i aus einer lognormal-Verteilung mit $a_i \sim logN(0, 0.25)$ gezogen. Personenparameter für den jeweils ersten Testzyklus wurden aus einer Standardnormalverteilung $\theta \sim N(0,1)$ gezogen. Für alle darauffolgenden Testzyklen t wurden die Personenparameter aus einer Normalverteilung $\theta \sim N(\mu_t, \sigma_t)$ gezogen, mit zufällig gezogenen Mittelwerten $\mu_t \in (-0.5, 0.0, 0.5)$ und Standardabweichungen $\sigma_t \in (0.7.\, 1.0, 1.3)$. Dieses Vorgehen simuliert den Umstand, dass sich

Testteilnehmer verschiedener Testzyklen in Bezug auf den Mittelwert und die Varianz ihrer Kompetenzverteilung unterscheiden können.

Zu Beantwortung von Fragestellung 1 dienten die auf den wahren Leichtigkeitsparameter d_i bedingten *MSE*s der Itemparameter d_i und a_i nach jedem Testzyklus als Evaluationskriterien. Zur Beurteilung der Qualität des Equatings (Fragestellungen 2 und 3) wurden der Anteil erfolgreicher Equatings (Anteil der Equatings, bei denen mindestens zwei Linkitems nach dem Test auf IPD übriggeblieben sind), der Anteil an als gedriftet identifizierten Items (sollte dem Alphafehlerniveau von .05 entsprechen), sowie der Fehler in den geschätzten Transformationskonstanten *A* und *B* innerhalb jedes Testzyklus (berechnet als Abweichung der geschätzten von den wahren Transformationskonstanten) genutzt.

Die Simulation wurde in R (R Core Team, 2021) unter Nutzung der Pakete mirt (Chalmers, 2012), mirtCAT (Chalmers, 2016), und equateIRT (Battauz, 2015) durchgeführt.

3.2.4 Ergebnisse

Wie bereits aufgrund der Ergebnisse von Beitrag 1 zu erwarten, nahm der *MSE* für den Diskriminations- und den Leichtigkeitsparameter im Verlauf der Testzyklen und mit zunehmender Stichprobengröße pro Testzyklus ab. Mit Bezug auf Fragestellung 1 zeigte sich, dass die Schwierigkeitsverteilung der Linkitems keinen substantiellen Einfluss auf die Messpräzision der Itemparameter hat, unabhängig von der genutzten Methode der Skalentransformation. Kleine Unterschiede ergaben sich lediglich für die bimodale Verteilung. Hier zeigte sich über den Verlauf der Testzyklen eine höhere Messpräzision in den Itemparametern für Items an den Rändern der Schwierigkeitsverteilung bei gleichzeitig niedrigerer Messpräzision für Items im mittleren Bereich im Vergleich zu den anderen Schwierigkeitsverteilungen der Linkitems (normal, gleichverteilt). Dieser Effekt zeigte sich vor allem bei sehr kleinen Stichroben (N = 50) und nahm mit zunehmender Stichprobengröße pro Testzyklus (N = 100, N = 300) ab.

Mit Bezug auf die Qualität des Equatings (Fragestellungen 2 und 3), ergab sich, dass über alle Replikationen der Anteil erfolgreicher Equatings bei 100% lag. Die Schwierigkeitsverteilung der Linkitems hatte keinen substantiellen Einfluss auf die anderen beiden Qualitätskriterien (Anteil als gedriftet detektierter Items, Fehler in den Transformationskonstanten). Bei normaler beziehungsweise gleichverteilter Schwierigkeitsverteilung der Linkitems wich der Anteil an gedrifteten Items unabhängig von der Skalentransformationsmethode kaum vom Alphafehlerniveau .05 ab (mit leichten Vorteilen für die Stocking-Lord Methode). Ausnahme war die Mean/Sigma-Methode, die bei sehr kleinen Stichproben (N = 50) generell zu deutlich niedrigeren Detektionsraten führte. Die Characteristic-Curve-Methoden (Haebara, Stocking-Lord) zeigten sich den Moment-Methoden (Mean/Mean, Mean/Sigma) in Bezug auf den Fehler in den geschätzten Transformationskonstanten *A* und *B* deutlich überlegen. Zwar wich der mittlere Fehler (Bias) in allen Bedingungen nicht wesentlich von Null ab, die Characteristic Curve Methoden zeigten

allerdings eine deutlich geringere Fehlervarianz. Vor allem bei sehr kleinen Stichprobengrößen (*N* = 50) zeigte auch hier die Mean/Sigma-Methode die schlechteste Performanz.

3.2.5 Diskussion

In Bezug auf die Präzision der Itemparameter lassen sich aus den Ergebnissen keine klaren Empfehlungen für die Schwierigkeitsverteilung der Linkitemparameter und die Skalentransformation ableiten. Da Itemparameter an den Rändern der Schwierigkeitsverteilung jedoch eher dazu neigen instabil zu sein, und die bimodale Verteilung keine substantiellen Vorteile hinsichtlich der Präzision der Itemparameter an den Rändern der Verteilung erbrachte, sollten eher gleich- beziehungsweise normalverteilte Schwierigkeitsverteilungen der Linkitems angestrebt werden. Auch hinsichtlich der Durchführbarkeit des Equatings ergaben sich keine Unterschiede zwischen den Bedingungen. Unter Verwendung der Momentmethoden (Mean/Mean, Mean/Sigma) traten jedoch recht extreme Fehlerwerte in den geschätzten Transformationskonstanten *A* und *B* auf. Unter den Characteristic-Curve-Methoden erwies sich die Stocking-Lord-Methode als geringfügig leistungsfähiger als die Haebara-Methode. Während also keine eindeutige Empfehlung hinsichtlich der Schwierigkeitsverteilung der Linkitemparameter ausgesprochen werden kann, erwies sich die Stocking-Lord-Methode als die leistungsfähigste Skalentransformationsmethode innerhalb des KKS.

3.3 Beitrag 3: Determinants of higher education teachers' intention to use technology-based exams

Zitation: Fink, A, Spoden, C. & Frey, A. (2022). Determinants of higher education teachers' intention to use technology-based exams. *International Journal of Educational Technology in Higher Education.* Manuskript eingereicht zur Publikation am 03.05.2022.

3.3.1 Einleitung

Eine erfolgreiche und flächendeckende Integration neuartiger Klausurensysteme in die alltägliche Prüfpraxis an Hochschulen hängt stark von der intendierten Nutzergruppe ab. Akzeptieren diese das neuartige System nicht, können sie ungeachtet derer Vorteile Implementationsprozesse erschweren oder sogar zum Scheitern bringen. Während sich die bisherige Forschung zur Implementation von E-Klausuren im Hochschulbereich eher auf die Akzeptanz von E-Klausuren seitens Studierender konzentriert hat (Maqableh et al., 2015; Terzis & Economides, 2011; Terzis et al., 2012; Zheng & Bender, 2019), besteht ein Bedarf an Studien, die explizit die Perspektive des akademischen Lehrpersonals auf E-Klausuren berücksichtigen (Bennett et al., 2017; Brady et al., 2019; Deeley, 2018). Insbesondere die Überzeugungen des Lehrpersonals, als für die Gestaltung und Implementation von Prüfungen verantwortliche Personengruppe, ist ein außerordentlich kritischer Faktor, der die erfolgreiche Implementation neuer Technologien in Lehrveranstaltungen

beeinflusst (z. B. Bennett et al., 2017; Brady et al., 2019; Nikou & Economides, 2018; Paiva et al., 2017). Darüber hinaus lassen sich von den wenigen Studien, die die Perspektiven der Hochschullehrenden auf E-Klausuren untersucht haben, noch weniger in einem klar definierten theoretischen Rahmen verorten (Brady et al., 2019). Übergeordnetes Ziel von Beitrag 3 ist es daher ein neuartiges Modell (Technology-based Exams Acceptance Model; kurz: TEAM) möglicher Einflussfaktoren auf die Intention zur Nutzung eines E-Klausurensystems aus Sicht von Hochschullehrenden herauszuarbeiten und empirisch zu prüfen. Das TEAM ist angelehnt an das Technology Acceptance Model (TAM; Davis, 1989) und dessen Erweiterungen (z. B. Terzis & Economides, 2011; Venkatesh et al., 2003) und umfasst die Variablen wahrgenommene Nützlichkeit, computerbezogene Selbstwirksamkeit, Computerängstlichkeit, Vorerfahrung im Sinne der Nutzung von digitalen Medien in Lehrveranstaltungen, günstige Rahmenbedingungen und subjektive Norm als Prädiktoren für die Intention zur Nutzung von E-Klausurensystemen (siehe Abbildung 2 im Vollbeitrag in Anhang C). Es soll zur Unterstützung zielorientierter und theoriebasierter Implementierungsprozesse nutzbar sein. Um einen solchen Einsatz zu rechtfertigen, wurden empirische Daten erhoben und mittels Strukturgleichungsmodellierung statistisch analysiert, um zu prüfen, ob die vorgeschlagene Modellstruktur zum tatsächlichen Antwortverhalten von Hochschullehrern passt. Darüber hinaus wird untersucht, ob bei der Implementation innovativer adaptiver E-Klausuren andere Bedingungen erfüllt sein müssen als bei konventionellen E-Klausuren.

3.3.2 Forschungsziele

Anhand empirischer Daten wird in dieser Studie die Fähigkeit des TEAMs, die Intention zur Nutzung von E-Klausuren zur summativen Leistungsüberprüfung seitens Hochschullehrender vorherzusagen, untersucht. Die Studie verfolgt dabei folgende drei Forschungsziele:

1. Formulierung des TEAMS
2. Untersuchung der Angemessenheit des TEAMS für Hochschullehrende
3. Statistischer Test der theoretisch abgeleiteten direkten und indirekten Effekte im TEAM
4. Untersuchung, ob sich die Ergebnisse zwischen konventionellen und adaptiven E-Klausuren unterscheiden.

3.3.3 Methode

Das TEAM wurde im Rahmen einer bundesweiten, hochschul- und fächerübergreifenden Online-Fragebogenstudie empirisch geprüft. Die Stichprobe wurde per E-Mail akquiriert. Die Teilnahme war freiwillig. Die nach der Datenbereinigung in die Analysen eingehende Stichprobe umfasste N = 992 Hochschullehrende (Geschlecht: 39.0 % weiblich, 61 % männlich; Alter: M = 44.29, SD = 11.87) aus 16 Bundeländern, 63 Universitäten und 35 Fachbereichen. Die Teilnehmerinnen und Teilnehmer wurden zufällig einer von zwei Gruppen zugewiesen. Gruppe 1 umfasste N_{ekl} = 494

Hochschullehrer (Geschlecht: 38.9 % weiblich, 61.1 % männlich; Alter: M = 44.30, SD = 11.95), die einen Fragebogen zu klassischen E-Klausuren beantworteten und Gruppe 2 umfasste N_{ad} = 498 Hochschullehrer (Geschlecht: 36.7 % weiblich, 63.3 % männlich; Alter: *M* = 44.28, *SD* = 11.80), die einen Fragebogen zu adaptiven E-Klausuren beantworteten. Es wurde eine Mischung aus bereits vorhandenen und neu entwickelten Skalen verwendet.

Das Modell wurde mittels Mehrgruppen-Strukturgleichungsmodellierung (MG-SEM) in Mplus 8.1 (Muthén & Muthén, 2018) getestet. Aufgrund des ordinalen Skalenniveaus der verwendeten Items wurde die Weighted Least Square Mean and Variance Adjusted (*WLSMV*) Schätzmethode genutzt. Zudem wurde die Mplus-Prozedur TYPE = COMPLEX verwendet, wodurch Modellfitstatistiken und Standardfehler für Fehlerabhängigkeiten aufgrund der genesteten Struktur der Daten (Hochschullehrer genestet in Hochschulen) angepasst werden. Nach Schätzung des Messmodells in beiden Gruppen und Analyse des Ladungsmusters, wurde eine Messinvarianzanalyse durchgeführt, um zu prüfen, ob die interessierenden Konstrukte in beiden Gruppen in vergleichbarer Weise gemessen wurden. In einem zweiten Schritt wurde das vollständige MG-SEM geschätzt und der Modellfit untersucht (Forschungsziel 2). Anschließend wurde die statistische Signifikanz der postulierten Zusammenhänge für jede Gruppe separat untersucht (Forschungsziel 3). Schließlich wurden die latenten Mittelwerte und Pfadkoeffizienten zwischen den beiden Gruppen mittels Wald-Test verglichen (Forschungsziel 4).

3.3.4 Ergebnisse

Bezogen auf Forschungsziel 2 erzielte das TEAM einen akzeptablen Modellfit: χ^2 = 5122.318, *df* = 3324, CFI = .947, TLI = .948, RMSEA = .033 (90% KI [.031, .035]), SRMR = .095. Durch das Modell konnten 78% (E-Klausuren) beziehungsweise 92% (adaptive E-Klausuren) der Varianz in der Intention zur Nutzung erklärt werden.

Bezüglich Forschungsziel 3 weisen die Ergebnisse die wahrgenommene Nützlichkeit als wichtigsten Prädiktor für die Intention zur Nutzung aus (p < .001). Vorerfahrung im Sinne der Nutzung von digitalen Medien in Lehrveranstaltungen hatte einen positiven Effekt auf die wahrgenommene Nützlichkeit (p < .001). Computerbezogene Selbstwirksamkeit wirkte nur indirekt über die Nutzungshäufigkeit digitaler Medien in Lehrveranstaltungen auf die wahrgenommene Nützlichkeit (p < .001). Computerängstlichkeit hatte einen negativen Effekt auf die computerbezogene Selbstwirksamkeit (p < .001). Darüber hinaus zeigten die Ergebnisse einen positiven Effekt der subjektiven Norm auf die wahrgenommene Nützlichkeit (p < .001) und die Intention zur Nutzung (p < .001). Günstige Rahmenbedingungen hatten keinen signifikanten Effekt auf die Intention zur Nutzung (siehe Abbildung 3 im vollständigen Beitrag im Anhang C).

Mit Bezug auf Forschungsziel 4 ergaben sich nur geringe Unterschiede zwischen den beiden Gruppen (bei gegebener Messinvarianz). So ergaben sich signifikante latente Mittelwertunterschiede nur für die beiden Faktoren wahrgenommene Nützlichkeit (d = -0.467, p <

.001) und Intention zur Nutzung (d = -0.274, p < .001), mit jeweils höheren Werten für Gruppe 1 (E-Klausuren) als für Gruppe 2 (adaptive E-Klausuren). Mit Blick auf die Pfadkoeffizienten zeigten die Ergebnisse, dass in Gruppe 1 (E-Klausuren) subjektive Norm einen signifikant größeren Effekt auf wahrgenommene Nützlichkeit (Δβ = 0.283; p = .007) und auf die Intention zur Nutzung (Δβ = 0.019; p = .008) hatte als in Gruppe 2 (adaptive E-Klausuren). Darüber hinaus hatte die wahrgenommene Nützlichkeit einen signifikant größeren Effekt auf die Intention zur Nutzung (Δβ = 0.097; p < .001) in Gruppe 1 (E-Klausuren).

3.3.5 Diskussion

Mit dem TEAM steht nun ein empirisch untersuchtes, hoch prädiktives Modell zur Erklärung der Intention zur Nutzung von (adaptiven) E-klausuren seitens Hochschulehrender zur Verfügung. Das Modell bietet eine fundierte theoretische Grundlage für die Optimierung von Implementierungsprozessen von E-Klausuren. Den Ergebnissen zufolge sollte der Fokus dabei darauf liegen, a) die computerbezogene Selbstwirksamkeitserwartung von Lehrenden zu fördern (zum Beispiel durch hochschuldidaktische Weiterbildungsangebote); b) die Hochschullehrenden zu ermutigen, verschiedene Arten von digitalen Medien in ihren Kursen auszuprobieren, damit sie sich mit ihnen vertraut machen und so im Allgemeinen Erfahrungen mit dem Einsatz von Technologien für Lehrzwecke sammeln können und schließlich c) durch eine geeignete Kommunikationsstrategie die Nützlichkeit von E-Klausuren zu verdeutlichen und sie zudem als Norm darzustellen.

3.4 Beitrag 4: Kriteriumsorientiertes adaptives Testen mit der KAT-HS-App

Zitation: Fink, A., Spoden, C., Frey, A. & Naumann, P. (2021). Kriteriumsorientiertes adaptives Testen mit der KAT-HS-App. *Diagnostica, 67*(2), 110–114. http://doi.org/10.1026/0012-1924/a000268

Bei Beitrag 4 handelt es sich um eine Softwareinformation, in der die im Dissertationsvorhaben entwickelte KAT-HS-App (KAT-HS = kriteriumsorientiertes adaptives Testen in der Hochschule) näher beschrieben ist. Ziel des Beitrages ist die Bekanntmachung und Nutzbarmachung der Software in der Breite. Die KAT-HS-App wurde entwickelt, um psychometrisch fundierte, digitale Hochschulklausuren gemäß dem Konzept von Spoden und Frey (2021) zu erstellen, kann allerdings auch in vielen weiteren Anwendungsbereichen zum Einsatz kommen. Von der Testkonstruktion über die Testadministration bis hin zur Testauswertung und Berichtlegung bündelt die App dafür eine Vielzahl von Methoden, für die üblicherweise unterschiedliche Softwarepakete eingesetzt werden müssen. Die Kernelemente der KAT-HS App sind: a) computerbasierte Testadministration; b) IRT-Skalierung; c) Methoden zur Überprüfung der psychometrischen Qualität des Tests; d) Kontinuierliche Kalibrierung bei wiederholten Anwendungen eines Tests; e) Kontrolle von Itempositionseffekten; f) computerisiertes adaptives Testen; und g) kriteriumsorientiertes Testen.

Zudem bietet sie eine grafische Nutzeroberfläche, eine intuitive Benutzerführung sowie für viele Anwendungsbereiche geeignete Voreinstellungen. Dies soll die Zugänglichkeit von IRT-basierten Methoden fächerübergreifend verbessern und gleichzeitig eine methodisch angemessene Nutzung dieser sicherstellen. Darüber hinaus kann mit der KAT-HS-App die in den Beiträgen 1 und 2 beschriebene und für viele Anwendungsbereiche attraktive kontinuierliche Kalibrierungsstrategie direkt genutzt werden. Die App basiert auf der Statistiksoftware R (R Core Team, 2021) und ist somit kostenlos und mit frei verfügbaren Quellcode erhältlich. Die App ist nach Registrierung über die Website https://kat-hs.uni-frankfurt.de/materialien/software/ für Forschung und Lehre kostenfrei verfügbar.

4 Diskussion

Im Folgenden werden die Ergebnisse der Arbeit zunächst kurz zusammengefasst (Kapitel 4.1). Anschließend werden die Limitationen aufgezeigt und ein Ausblick für zukünftige Forschungsaktivitäten gegeben (Kapitel 4.2). Schließlich wird der Erkenntnisgewinn der Arbeit herausgestellt und hinsichtlich seiner theoretischen und praktischen Relevanz bewertet (Kapitel 4.3).

4.1 Zusammenfassung der Ergebnisse

Ausgangsunkt der hier vorliegenden Arbeit war, dass CAT vor allem durch seine hohe Messeffizienz sowie Vorteile, die aus der computerisierten Testadministration resultieren, zwar das Potential besitzt die Qualität von Hochschulklausuren deutlich zu steigern, jedoch nicht ohne weiteres auf den Bereich von Hochschulklausuren angewendet werden kann. Übergeordnetes Ziel dieser Arbeit war es daher, zentrale Herausforderungen auf psychometrischer, personaler sowie technischer Ebene, die im Zusammenhang mit der Implementation kriteriumsorientierter adaptiver Hochschulklausuren stehen, zu überwinden.

Die ersten beiden Fragestellungen bezogen sich dabei auf die psychometrischen Herausforderungen. Da aufgrund der üblicherweise an Hochschulen vorzufindenden Rahmenbedingungen (z. B. vergleichsweise kleine Kalibrierungsstichproben, begrenzte Ressourcen für die Itementwicklung) eine separate Kalibrierungsstudie vor der eigentlichen Testanwendung für adaptive Hochschulklausuren nur schwer realisierbar ist, lautete die erste Fragestellung, wie die Kalibrierung des Itempools für kriteriumsorientierte, adaptive Hochschulklausuren im laufenden Lehrbetrieb erfolgen kann (Fragestellung 1). Als Antwort hierauf wurde in Beitrag 1 eine neuartige Kalibrierungsstrategie vorgestellt und hinsichtlich ihrer psychometrischen Eigenschaften untersucht. Es konnte gezeigt werden, dass die KKS selbst für sehr kleine Stichproben eine geeignete Methode darstellt, den Itempool über mehrere Testanwendungen zu kalibrieren, sodass über Testanwendungen hinweg die Tests immer präziser messen, defizitäre Items identifiziert werden können und die Skala konstant gehalten wird. Eine bedeutende Hürde für den Einsatz von CAT für Hochschulklausuren kann somit überwunden werden.

Um Vergleichbarkeit der Ergebnisse über mehrere Klausurzeitpunkte und somit Fairness nicht nur innerhalb einer Kohorte, sondern auch zwischen aufeinanderfolgenden Kohorten zu gewährleisten, sollten die Ergebnisse einzelner Klausurdurchläufe statistisch verbunden sein. Die KKS nutzt daher Equating-Methoden, um die Skala über Testanwendungen hinweg konstant zu halten. Um eine bestmögliche Vergleichbarkeit der Ergebnisse bei gleichzeitig bestmöglicher Effizienz der KKS gewährleisten zu können, widmete sich der zweite Beitrag daher der Frage, wie die Equating-Prozedur der KKS konfiguriert werden sollte, um eben dieses Ziel zu erreichen

(Fragestellung 2). Hierfür wurden unter anderem die Schwierigkeitsverteilung der genutzten Linkitems sowie die genutzte Skalentransformationsmethode variiert. Unter verschiedenen Konfigurationen zeigte sich die KKS in der Lage, die Skala über mehrere Testzyklen hinweg konstant zu halten. Normal- beziehungsweise gleichverteile Schwierigkeitsverteilungen der Linkitems sowie die Stocking-Lord-Skalentransformationsmethode erzielten hierbei die besten Ergebnisse.

Neben psychometrischen Hürden, die es zu überwinden gilt, spielt für die erfolgreiche Implementation adaptiver Klausurensysteme auf personaler Ebene vor allem die Akzeptanz seitens der Lehrenden eine entscheidende Rolle. Beitrag 3 widmete sich daher der Fragestellung, welche Faktoren die Intention zur Nutzung von kriteriumsorientierten, adaptiven Hochschulklausuren seitens Hochschullehrender beeinflussen (Fragestellung 3). Zur Beantwortung der Frage wurde auf Basis bestehender Modelle der Technologieakzeptanz das Technology-based Exams Acceptance Model (TEAM) vorgeschlagen und empirisch geprüft. Die Ergebnisse identifizierten die wahrgenommene Nützlichkeit als Haupteinflussfaktor auf die Nutzungsintention. Computerbezogene Selbstwirksamkeit sowie die Computerängstlichkeit hatten indirekte Effekte auf die wahrgenommene Nützlichkeit, die durch vorherige Erfahrungen mit dem Einsatz digitaler Medien für Lehrzwecke mediiert wurden. Außerdem stellte sich die subjektive Norm als ein wichtiger Einflussfaktor auf die Intention zur Nutzung von E-Klausuren heraus. Das TEAM stellt eine solide, empirisch gestützte theoretische Basis zur Unterstützung von Implementationsprozessen von (adaptiven) E-Klausuren an Hochschulen dar.

Um eine Anwendung kriteriumsorientierter, adaptiver Hochschulklausuren in der breite realistisch zu machen, wurde als notwendige Bedingung auf technischer Ebene schließlich die Bereitstellung einer geeigneten Software konstatiert. Als Antwort auf die Frage, wie die Konstruktion, Administration und Auswertung kriteriumsorientierter, adaptiver Hochschulklausuren softwaretechnisch auch für Personen mit geringeren psychometrischen Vorkenntnissen ermöglicht werden kann (Fragestellung 4), wurde im Rahmen des Dissertationsvorhabens die KAT-HS-App entwickelt, die in Beitrag 4 beschrieben ist. Sie bildet damit den logischen Abschluss dieser Dissertation, indem sie die zentral notwendigen IRT-basierten Verfahren in einem Softwarepaket bündelt und ihre Anwendung in der Breite ermöglicht. So ist mit ihr auch die in den Beiträgen 1 und 2 beschriebene KKS direkt umsetzbar. Durch das Bereitstellen einer grafischen Nutzeroberfläche, geeigneten Voreinstellungen, Vorlagen und einem Benutzerhandbuch ermöglicht sie einer breiten Nutzerschaft die Anwendung adaptiver und nicht-adaptiver IRT-basierter E-Klausuren.

4.2 Limitationen und Ausblick

Auch wenn jeder der Einzelbeiträge der hier vorliegenden Dissertation einen substantiellen Beitrag zur Überwindung zentraler Herausforderungen bei der Implementation adaptiver Hochschulklausuren auf psychometrischer, technischer und personaler Ebene liefert, existieren

weitere Aspekte, die nicht im Rahmen dieser Arbeit betrachtet werden konnten und somit Untersuchungsgegenstand zukünftiger Studien sein sollten.

Ein wichtiger Aspekt aus psychometrischer Sicht, der in den Beiträgen 1 und 2 nicht berücksichtigt wurde, sind Itempositionseffekte (IPE). Als IPE wird die systematische Variation von statistischen Eigenschaften von Items in Abhängigkeit ihrer Darbietungsposition in einem Test bezeichnet (z. B. Frey et al., 2017). Empirisch zeigt sich typischerweise, dass der Anteil korrekter Antworten auf ein bestimmtes Item abnimmt, je weiter hinten im Test es präsentiert wird, was wiederum gleichbedeutend mit einer Zunahme der Itemschwierigkeit ist (z. B. Albano, 2013; Debeer et al., 2014; Nagy et al., 2019; Wu et al., 2019). Das führt wiederum dazu, dass eine korrekte Antwort auf ein und dasselbe Item in Abhängigkeit zu dessen Darbietungsposition im Test zu einer unterschiedlichen Kompetenzschätzung führen würde. Somit können sich IPEs bei Nichtbeachtung negativ auf die Validität der Testwertinterpretationen auswirken. So ein Muster ist vor allem für CAT ein Problem, da die individualisierte Testzusammenstellung dazu führt, das verschiedene Personen verschiedene Items an verschiedenen Positionen bearbeiten, während für alle Positionen die gleichen Itemparameter für die Itemauswahl und die Personenparameterschätzung verwendet werden. Bei Vorhandensein von IPEs kann dies wiederum zu systematischen Verzerrungen bei der Personenparameterschätzung führen und somit die Validität der Testwertinterpretationen einschränken (Frey & Fink, im Druck), was bezogen auf Hochschulklausuren wiederum ein Problem im Sinne des Gleichbehandlungsgrundsatzes darstellt. Daher ist es wichtig IPEs bereits bei der Kalibrierung von CAT-Itempools zu berücksichtigen (Frey et al., 2017). Aufbauend auf diesen Erkenntnissen schlugen Frey und Fink (im Druck) eine balancierte Variante der KKS vor und konnten in ihrer Studie zeigen, dass diese in der Lage ist für Itempositionseffekte zu kontrollieren. Daher ist die balancierte KKS auch bereits in die KAT-HS-App integriert.

Darüber hinaus können Potentiale von CAT nur dann vollumfänglich ausgeschöpft werden, wenn das Klausurensystem die gegebenen Antworten automatisch bewerten kann (engl. automated scoring; z. B. Yan et al., 2020). Somit eignen sich vor allem geschlossene Itemformate sowie halboffene Items mit Kurzantworten für CAT. Offene Antwortformate können zwar ebenfalls in adaptive Klausuren eingebunden werden, sie können allerdings nur auf Grundlage der vorläufigen Kompetenzschätzung ausgewählt, und die gegebenen Antworten gespeichert werden. Für die adaptive Itemauswahl und Kompetenzschätzung während des Tests können die Antworten auf die offenen Items nicht genutzt werden (Fink et al., 2021). Aktuelle Methoden aus dem Bereich des Natural Language Processings (NLP; z. B. Cahill & Evanini, 2020) bieten hier vielversprechende Lösungsansätze, wie auch offene Antwortformate automatisch ausgewertet und somit vollumfänglich für CAT verwendet werden können. Zukünftige Arbeiten sollten sich daher damit beschäftigen, wie solche Methoden auch für den Bereich von Hochschulklausuren in sowohl psychometrisch als auch prüfungsrechtlich angemessener Weise eingesetzt werden können, um die fächerübergreifende Nutzbarkeit adaptiver Hochschulklausuren noch weiter zu erhöhen.

Auf personaler Ebene wurde in dieser Dissertation lediglich die Akzeptanz seitens Hochschulehrender in den Blick genommen. Auch wenn diese eine entscheidende Stellschraube für die erfolgreiche Implementation adaptiver Hochschulklausuren darstellt, dürfen auch die Lernenden nicht außer Acht gelassen werden (z. B. Romeu Fontanillas et al., 2016). Obwohl die hier präsentierte, innovative Art von Hochschulklausuren zahlreiche Vorteile sowohl aus psychometrischer als auch hochschuldidaktischer Sicht bietet, ist es wichtig zu untersuchen, ob eine Integration von adaptiven Hochschulklausuren in den regulären Prüfbetrieb von Hochschulen mit Widerstand seitens der Studierenden verbunden ist, der aus negativem Affekt gegenüber der Anwendung von CAT für Hochschulklausuren resultiert. Ein zentraler Aspekt an dieser Stelle ist, dass bei CAT das nachträgliche Überarbeiten von bereits abgegeben Antworten (sog. Itemrevision) üblicherweise nicht möglich ist, da dies die Effizienz des adaptiven Algorithmus verringert und zu einer Verzerrung der Personenparameterschätzung führen kann (z. B. Stocking, 1997; Wang et al., 2017). Die fehlende Möglichkeit zur Itemrevision kann zu einem niedrigerem Kontrollerleben seitens der Studierenden im Vergleich zu herkömmlichen Klausuren führen (Naumann & Fink, 2021). Nach der Kontrolle-Wert-Theorie der Leistungsemotionen (Pekrun et al., 2007; Pekrun & Perry, 2014) kann ein solcher Kontrollverlust wiederum in negativen Emotionen wie beispielsweise Testangst resultieren. Neuartige CAT-Verfahren, die auch Itemrevision erlauben, haben das Potential, diesen negativen Effekten entgegenzuwirken (Esmaeili Bijarsari et al., 2019; Naumann & Frey, 2021; Olea et al., 2000) und sollten daher zukünftig stärker in den Fokus gerückt und für Hochschulklausuren nutzbar gemacht werden. So kann von den Vorteilen von CAT profitiert werden, ohne dass ein Mangel an subjektivem Kontrollerleben negative emotionale Auswirkungen auf die Studierenden hat.

Aus technischer Sicht bringen alle der oben angesprochenen Aspekte mit sich, dass die Entwicklung der KAT-HS-App nicht an diesem Punkt stoppt. Vielmehr ist es notwendig sie kontinuierlich auf Grundlage neuester Erkenntnisse und Methoden aus den Bereichen Psychometrie und Educational Measurement, sowie den realen hochschuldidaktischen Anforderungen weiterzuentwickeln. Neben den oben bereits genannten Punkten sehen künftige Entwicklungsschritte beispielsweise vor den Funktionsumfang der KAT-HS-App um ordinale IRT-Modelle (z. B. Muraki & Muraki, 2016), Personen-Fit Analysen (z. B. Emons et al., 2005) und Möglichkeiten zum automatisierten Feedback mittels Natural Language Generation (z. B. Gatt & Krahmer, 2018) zu erweitern. Die Weiterentwicklung der App soll vor allem durch das Bilden einer aktiven Nutzercommunity vorangetrieben werden.

Abschließend ist noch anzumerken, dass auch die in dieser Dissertation bislang nicht behandelte rechtliche Ebene eine zentrale Herausforderung bei der Implementation kriteriumsorientierter, adaptiver Klausuren darstellt (z. B. Frey et al., 2021). Bestehende prüfungsrechtliche Grundsätze sind mit den technischen Entwicklungen moderner, computerbasierter Testverfahren in Einklang zu bringen. Diesbezügliche prüfungsrechtliche

Diskussionen, wie sie bei Fink (2021) oder Frey et al. (2020) zu finden sind, lassen aber darauf schließen, dass adaptive Klausuren keinen prüfungsrechtlichen Grundsätzen, so wie sie in deutschen Hochschulen gehandhabt werden, widersprechen. Da solche Klausuren an deutschen Hochschulen bislang nicht im Regelbetrieb eingesetzt werden, steht eine abschließende rechtliche Einschätzung vor Gericht allerdings noch aus.

4.3 Fazit

Die fortschreitende Flexibilisierung und Individualisierung von Studiengängen und die Forderung nach konsequent kompetenzorientierter Lehre auf der einen, sowie die weitreichenden Einflüsse der Digitalisierung auf die Hochschullehre auf der anderen Seite bringen mit sich, dass auch Hochschulklausuren neu gedacht werden müssen. Insbesondere adaptive E-Klausuren bieten hier ein noch nicht systematisch genutztes Potential, Einschränkungen, die klassische Klausuren mit sich bringen, zu überwinden und Hochschulklausuren flexibler und individualisierter und dabei gleichzeitig messpräziser und schließlich fairer zu machen (Spoden & Frey, 2021). Um vom Versprechen eines Qualitätssprungs in diesem zentralen Bereich der Hochschullehre profitieren zu können, gilt es allerdings zunächst zentrale Herausforderungen bei der Implementation adaptiver E-Klausuren zu überwinden. Die Ergebnisse dieser Arbeit liefern hierfür zentrale Grundlagen.

So ermöglicht es die KKS, orientiert an den verfügbaren Ressourcen für die Testentwicklung, einen CAT-Itempool über mehrere Testzeitpunkte sukzessive aufzubauen, wobei die Skala über Testzeitpunkte hinweg konstant gehalten wird. Die sonst üblicherweise notwendige Kalibrierungsstudie entfällt somit. Eine solche Methode ist nicht nur für Hochschulklausuren von praktischer Relevanz, sondern für alle potentiellen Einsatzgebiete von CAT, in denen aus Ermangelung an verfügbaren Ressourcen für die Testentwicklung CAT bisher üblicherweise nicht zum Einsatz kommt (z. B. psychologische Tests für die Personalauswahl, klinische Diagnostik). Aus den Ergebnissen aus den Beiträgen 1 und 2 lassen sich zudem praktische Empfehlungen für die Konfiguration des Algorithmus ableiten. Die KKS kann in ihrer balancierten Erweiterung (Frey & Fink, im Druck) mit der KAT-HS-App direkt umgesetzt werden.

Darüber hinaus wurden mit dem TEAM zentrale Einflussfaktoren auf die Intention zur Nutzung von adaptiven und nicht-adaptiven E-Klausuren seitens Hochschullehrender identifiziert. Den Studienergebnissen zufolge sollte der Fokus im Implementationsprozess insbesondere darauf gelegt werden, a) die computerbezogene Selbstwirksamkeit der Lehrenden zu fördern (z. B. durch Hochschuldidaktische Weiterbildungen und Schulungsmaterialien); b) die Lehrenden zu ermutigen, verschiedene Arten digitaler Medien für Lehrzwecke zu nutzen, damit sie sich mit ihnen vertraut machen und so Erfahrungen mit dem Einsatz von Bildungstechnologien im Allgemeinen sammeln können, und c) E-Klausuren durch eine geeignete Kommunikationsstrategie hochschulweit zu bewerben. Das Modell sowie die mit der Studie publizierten Skalen können zudem dafür genutzt werden, die Wirksamkeit von Interventionsmaßnahmen zu evaluieren, die auf diese Ziele abzielen.

Schließlich bietet die KAT-HS-App eine benutzerfreundliche All-in-one-Lösung für die Konstruktion, Administration und Auswertung adaptiver und nicht-adaptiver Hochschulklausuren, welche durch eine offene GNU General Public License (GNU GPL) nicht in Abhängigkeitsverhältnissen mündet, und an die jeweiligen Bedürfnisse der Hochschule, der Fachbereiche oder der Modulverantwortlichen anpassbar ist. Einer solch offenen Software ist es zudem möglich, den rasant voranschreitenden Entwicklungen in den Bereichen Psychometrie und pädagogisch-psychologischer Diagnostik standzuhalten, in dem ihre Weiterentwicklung gemäß neuester wissenschaftlicher Erkenntnisse durch den Aufbau und die Förderung dezentraler, kooperativer Entwicklungsprozesse vorangetrieben wird (Frey et al., 2021).

In der Zusammenschau liegen mit dieser Dissertation also direkt einsetzbare Bausteine für die Überwindung zentraler Herausforderungen bei der zielgerichteten Implementation kriteriumsorientierter, adaptiver E-Klausuren vor. In Kombination mit niedrigschwelligen hochschuldidaktischen Weiterbildungsangeboten für Prüferinnen und Prüfer verschiedener Fachbereiche, sowie dem Aufbau technischer und prüfungsdidaktischer Supportstrukturen, ist der Weg geebnet für moderne Hochschulklausuren, die den Anforderungen individualisierter Lehre gerecht werden können und gleichzeitig valide Schlüsse auf das Erreichen kompetenzorientierter Lehrziele zulassen.

Literatur

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement, 50*, 408–426. https://doi.org/ 10.1111/jedm.12026

Asseburg, R. (2011). *Leistungsbereitschaft in Testsituationen. Motivation zur Bearbeitung adaptiver und nicht-adaptiver Leistungstests.* Marburg: Tectum.

Babcock, B. & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, *1*, 1–18. https://doi.org/10.7333/1212-0101001

Bandtel, M., Baume, M., Brinkmann, E., Bedenlier, S. Budde, J., Eugster, B., Andrea, G., Halbherr, T., Persike, M., Rampelt, F., Reinmann, G., Sari, Z. & Schulz, A. (Hrsg.) (2021). *Digitale Prüfungen in der Hochschule: Whitepaper einer Community Working Group aus Deutschland, Österreich und der Schweiz.* Hochschulforum Digitalisierung. https://hochschulforumdigitalisierung.de/sites/default/files/dateien/HFD_Whitepaper_Digitale_Pruefungen_Hochschule.pdf

Battauz, M. (2015). equateIRT: an R package for IRT test equating. *Journal of Statistical Software, 68*, 1–22. https://doi.org/10.18637/jss.v068.i07

Bedenlier, S., Bandtel, M., Boom, K.-D., Gerl, S., Haldbherr, T., Hebel, A.-L., Jeremias, X., Kehr, H., Mecklenburg, L., Mersch, A., Molter, K., Paffenholz, A., Reinmann, G., Riebe, K. & van Treeck, T. (2021). Prüfungen aus Perspektive der Prüfungsdidaktik. In M. Bandtel, M. Baume, E. Brinkmann, S. Bedenlier, J. Budde, B. Eugster, G. Andrea, T. Halbherr, M. Persike, F. Rampelt, G. Reinmann, Z. Sari & A. Schulz (Hrsg.). *Digitale Prüfungen in der Hochschule: Whitepaper einer Community Working Group aus Deutschland, Österreich und der Schweiz* (S. 30–42). Hochschulforum Digitalisierung.

Bennett, S., Dawson, P., Bearman, M., Molloy, E. & Boud, D. (2017). How technology shapes assessment design: Findings from a study of university teachers. *British Journal of Educational Technology, 48,* 672–682. *https://doi.org/10.1111/bjet.12439*

Bergstrom, B. A., Lunz, M. E. & Gershon, R. C. (1992). Altering the level of difficulty in computeradaptive testing. *Applied Measurement in Education, 5,* 137–149. https://doi.org/10.1207/s15324818ame0502_4

Betz, N. E. (1975). New types of information and psychological implications. In D. J. Weiss (Ed.), *Computerized adaptive trait measurement: Problems and Prospects (Research Report 75-5)* (pp. 32–43). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Betz, N. E. & Weiss, D. J. (1976a). *Effects of immediate knowledge of results and adaptive testing on ability test performance (Research Report 76-3).* Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Betz, N. E. & Weiss, D. J. (1976b). *Psychological effect of immediate knowledge of results and adaptive ability testing (Research Report 76-4)*. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364. https://doi.org/10.1007/bf00138871

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444. https://doi.org/10.1177%2F014662168200600405

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y. & Bosker, R. J. (2015). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PloS one*, *10*(12), e0143616. https://doi.org/10.1371/journal.pone.0143616

Born, S. & Fink, A. (2021). Etablierung einer stabilen Berichtsmetrik bei Hochschulklausuren. In C. Spoden & A. Frey (Hrsg.), *Psychometrisch fundierte E-Klausuren für die Hochschule* (S. 65–72). Lengerich: Pabst Science Publishers.

Born, S. & Frey, A. (2017). Heuristic constraint management methods in multidimensional adaptive testing. *Educational and Psychological Measurement*, *77*, 241–262. https://doi.org/10.1177/0013164416643744

Born, S. & Spoden, C. (2021). Skalierung von Hochschulklausuren. In C. Spoden & A. Frey (Hrsg.), *Psychometrisch fundierte E-Klausuren für die Hochschule* (S. 27–38). Lengerich: Pabst Science Publishers.

Brady, M., Devitt, A. & Kiersey, R. A. (2019). Academic staff perspective on technology for assessment (TfA) in higher education: A systematic literature review. *British Journal of Educational Technology, 50*, 3080–3098. https://doi.org/10.1111/bjet.12742

Cahill, A. & Evanini, K. (2020). Natural language processing for writing and speaking. In D. Yan, A. A. Rupp & P. W. Foltz (Eds.) *Handbook of automated scoring: Theory into practice* (pp. 69–92). Boca Raton, FL: Chapman & Hall/CRC.

Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1–29. https://doi.org/10.18637/jss.v071.i05

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71,* 1–39. https://doi.org/10.18637/jss.v071.i05

Cheng, P. E. & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement*, *24*, 257–265. https://doi.org/10.1177%2F01466210022031723

Chua, Y. P. & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior, 29,* 1889–1895. https://doi.org/10.1016/j.chb.2013.03.008

Cronbach, L. J. & Gleser, G.C. (1965). *Psychological Tests and Personnel Decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user accpetance of information technology. *MIS Quarterly, 13(3),* 319–339. https://doi.org/10.2307/249008

de Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd ed.). New York, NY: The Guilford Press.

Debeer, D., Buchholz, J., Hartig, J. & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39*, 502–523. https://doi.org/10.3102%2F1076998614558485

Dolan, R. P. & Burling, K. S. (2017). Computer-based testing in higher education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (2nd ed., pp. 370–384). New York, NY: Routledge. https://doi.org/10.4324/9781315709307.ch24

Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing*. Enschede: Print Partners Ipskamp.

Eggen, T. J. H. M. & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30,* 379–393. https://doi.org/10.1177/0146621606288890

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Emons, W. H., Sijtsma, K. & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods, 10*(1), 101–119. https://doi.org/10.1037/1082-989X.10.1.101

Esmaeili Bijarsari, S., Frey, A., Spoden, C., Born, S. & Fink, A. (2019, Februar). Emotionale Effekte von Itemreview in Hochschulklausuren [Konferenzbeitrag]. 7. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF), Köln, Germany.

Fink, A. (2021). Rechtliche Aspekte computerbasierter und adaptiver Hochschulklausuren. In C. Spoden & A. Frey (Hrsg.), *Psychometrisch fundierte E-Klausuren für die Hochschule* (S. 84–93) Lengerich: Pabst Science Publishers.

Fink, A., Naumann, P. & König, C. (2021). Computerisierte adaptive Klausuren im Psychologiestudium. *Psychologische Rundschau*, *72*(2), 125-127. http://doi.org/10.1026/0033-3042/a00053

Frey, A. (2006). *Validitätssteigerungen durch adaptives Testen*. Frankfurt am Main: Peter Lang.

Frey, A. (2020). Computerisiertes adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 501–524). Springer. https://doi.org/10.1007/978-3-662-61532-4_20

Frey, A. (2021). Individualisierung von Hochschulklausuren durch adaptives Testen. In C. Spoden & A. Frey (Hrsg.), *Psychometrisch fundierte E-Klausuren für die Hochschule* (S. 55–64). Lengerich: Pabst Science Publishers.

Frey, A., Bernhardt, R. & Born, S. (2017). Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests. *Diagnostica, 63,* 167–178. https://doi.org/10.1026/0012-1924/a000173

Frey, A. & Fink, A. (im Druck). Controlling for item position effects when adaptive testing is used in Large-Scale Assessments. In L. Khorramdel, M. von Davier & K. Yamamoto (Eds.), *Innovative computer-based international large-scale assessments – foundations, methodologies and quality assurance procedures*. Springer.

Frey, A. & Hartig, J. (2022). Kompetenzdiagnostik. In M. Harring, C. Rohlfs & M. Gläser-Zikuda (Hrsg.), *Handbuch Schulpädagogik* (2. Aufl., S. 928–937). Münster: Waxmann.

Frey, A., Hartig, J. & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. *Diagnostica, 55*, 20–28. https://doi.org/10.1026/0012-1924.55.1.20

Frey, A., Spoden, C. & Born, S. (2020). Construction of psychometrically sound written university exams. *Psychological Test and Assessment Modeling, 65,* 472–486. https://www.psychologie-aktuell.com/journale/psychological-test-and-assessment-modeling/currently-available/inhaltlesen/psychological-test-and-assessment-modeling-2020-4.html

Frey, A., Spoden, C., Fink, A. & Born, S. (2020). Kompetenzorientierte individualisierte Hochschulklausuren und deren prüfungsrechtliche Einordnung. *eleed, 13.* urn:nbn:de:0009-5-51197

Frey, A., Spoden, C. & Schultze, M. (2021). Die Zukunft der Hochschulklausuren hat bereits begonnen. *Psychologische Rundschau, 72*, 113–116. https://doi.org/10.1026/0033-3042/a000528

Gatt, A. & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research, 61*(2), 65–170. https://doi.org/10.5555/3241691.3241693

Goldstein, H. (1983). Measuring changes in educational attainment over time: problems and possibilities. *Journal of Educational Measurement, 20*, 369–377. https://doi.org/10.1111/j.1745-3984.1983.tb00214.x

Gu, L., Ling, G., Liu, O. L., Yang, Z., Li, G., Kardanova, E. & Loyalka, P. (2020). Examining mode effects for an adapted Chinese critical thinking assessment. *Assessment & Evaluation in Higher Education, 46*, 870–893. https://doi.org/10.1080/02602938.2020.1836121

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144–149. https://doi.org/10.4992/psycholres1954.22.144

Häfer, J. & Matthé, F. (2016). Ein nach vorne offener Prozess: E-Assessments an Hochschulen. *Forschung & Lehre*, *3*, 195–197.

Hartig J., Frey A. & Jude N. (2020) Validität von Testwertinterpretationen. In Moosbrugger H., Kelava A. (Hrsg.) *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 529–545). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-61532-4_21

Herzberg, P. Y. & Frey, A. (2011). Kriteriumsorientierte Diagnostik. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Methoden der psychologischen Diagnostik. Enzyklopädie der Psychologie, B/II/2* (S. 281–324). Göttingen: Hogrefe.

Kim, S. (2006). A comparative study of IRT fixed parameters calibration methods. *Journal of Educational Measurement, 43,* 355–381. https://doi.org/10.1111/j.1745-3984.2006.00021.x

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, *52*, 876–903.

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer. http://doi.org/10.1007/978-1-4939-0317-7_10

Kong, X. J., Wise, S. L., Harmes, J. C. & Yang, S. (2006, April). *Motivational effects of praise in response-time based feedback: A follow-up study of the effort-monitoring CBT*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Leroux, A. J., Lopez, M., Hembry, I. & Dodd, B. G. (2013). A comparison of exposure control procedures in CATs using the 3PL model. *Educational and Psychological Measurement, 73*, 857–874. https://doi.org/10.1177%2F0013164413486802

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, MI: Lawrence Erlbaum Associates.

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the rasch model. *Journal of Educational Measurement, 17,* 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Makransky, G. & Glas, C. A. W. (2010). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology, 11*, 1–20. http://www.jattjournal.net/index.php/atp/article/view/48350

Maqableh, M., Masa'deh, R. & Mohammed, A. B. (2015). The acceptance and use of computer based assessment in higher education. *Journal of Software Engineering and Applications, 8*, 557–574. https://doi.org/10.4236/jsea.2015.810053

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

McBride, J. R. (1976). Bandwidth, fidelity, and adaptive tests. In. T. J. McConell, Jr. (Ed.) C*AT/C 2 1975: The second conference on computer-assisted test construction* (pp. 81 – 98). Atlanta GA: Atlanta Public Schools.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177–195. https://doi.org/10.1007/BF02293979

Moosbrugger, H. & Goldhammer, F. (2007). *FAKT-II. Frankfurter Adaptiver Konzentrationsleistungs-Test*. Bern: Huber.

Muraki, E. & Muraik, M. (2016). Generalized partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Models* (S. 127–137). Boca Raton, FL: CRC Press Taylor & Francis Group.

Muthén, L. K. & Muthén, B. O. (2018). *Mplus User's Guide, 8th Edn.* Los Angeles, CA: Muthén & Muthén.

Nagy, G., Nagengast, B., Frey, A., Becker, M. & Rose, N. (2019). A multilevel study of position effects in PISA achievement tests: Student- and school-level predictors in the German tracked school system. *Assessment in Education: Principles, Policy & Practice, 26,* 422–443. https://doi.org/10.1080/0969594X.2018.1449100

Nardi, A. & Ranieri, M. (2018). Comparing paper-based and electronic multiple-choice examinations with personal devices: Impact in students' performance, self-efficacy and satisfaction. *British Journal of Educational Technology, 50,* 1495–1506. https://doi.org/10.1111/bjet.12644

Naumann, P. & Fink, A. (2021). Potentielle Hürden bei der Nutzung psychometrisch fundierter Hochschulklausuren. In C. Spoden & A. Frey (Hrsg.), *Psychometrisch fundierte E-Klausuren für die Hochschule* (S. 100–105) Lengerich: Pabst Science Publishers.

Naumann, P. & Frey, A. (2021, September). Emotionale Effekte von computerisierten adaptiven Hochschulklausuren [Posterpräsentation]. 15. Tagung der Fachgruppe Methoden und Evaluation der Deutschen Gesellschaft für Psychologie (DGPs), Mannheim, Germany.

Nikou, S. A. & Economides, A. A. (2016). The impact of paper-based, computer-based and mobile-based self-assessment on students' science motivation and achievement. *Computers in Human Behavior, 55*, 1241–1248. https://doi.org/10.1016/j.chb.2015.09.025

Nikou, S. A. & Economides, A. A. (2018). Mobile-based assessment: A literature review of publications in major referred journals from 2009 to 2018. *Computers & Education, 125,* 101–119. *https://doi.org/10.1016/j.compedu.2018.06.006*

Olea, J., Revuelta, J., Ximenez, M. C. & Abad, F. J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicológica, 21*, 157–173. https://www.uv.es/revispsi/articulos1y2.00/olea.pdf

Ones, D. S. & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior, 17,* 609–626. https://doi.org/10.1002/(SICI)1099-1379(199611)17:6%3C609::AID-JOB1828%3E3.0.CO;2-K

Ortner, T. M. & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*, *27*, 157–163. https://doi.org/10.1027/1015-5759/a000062

Ortner, T. M., Weißkopf, E. & Koch, T. (2014). I will probably fail: Higher ability students' motivational experiences during adaptive achievement testing. *European Journal of Psychological Assessment, 30*, 48-56. https://doi.org/10.1027/1015-5759/a000168

Paiva, J., Morais, C., Costa, L. & Pinheiro, A. (2017). The shift from "e-learning" to "learning": Invisible technology and the dropping of the "e". *British Journal of Educational Technology, 47,* 226–238. https://doi.org/10.1111/bjet.12242

Pekrun, R., Frenzel, A. C., Goetz, T. & Perry, R. P. (2007). The control-value theory of achievement emotions: An integrative approach to emotions in education. In P. A. Schutz & R. Pekrun (Hrsg.), *Emotion in education* (S. 13–36). New York, NY: Academic Press. https://doi.org/10.1016/B978-0-12-372545-5.X5000-X

Pekrun, R. & Perry, R. P. (2014). Control-value theory of achievement emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 120–141). Taylor & Francis.

Pine, S. M., Church, A. T., Gialluca, K. A. & Weiss, D. J. (1979). *Effects of computerized adaptive testing on black and white students (Research Report 79-2).* Minneapolis, MN: Department of Psychology, University of Minnesota.

Ponsoda, V., Olea, J., Rodriguez, M. S. & Revuelta, J. (1999). The effects of test difficulty manipulationin computerized adaptive testing and self-adapted testing. *Applied Measurement in Education, 12*, 167–184. https://doi.org/10.1207/s15324818ame1202_4

R Core Team (2021). *R: A Language and Environment for Statistical Computing [Software].* Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

Reckase, M. D. (2016). Logistic Multidimensional Models. In W. J. van der Linden (Ed.). *Handbook of item response theory, Volume one: Models* (pp. 189–209). Boca Raton, FL: Chapman & Hall/CRC.

Rolim, C. & Isaias, P. (2019). Examining the use of e-assessment in higher education: Teachers and students' viewpoints. *British Journal of Educational Technology, 50,* 1785–1800. https://doi.org/10.1111/bjet.12669

Romeu Fontanillas, T., Romero Carbonell, M. & Guitert Catasús, M. (2016). E-assessment process: giving a voice to online learners. *International Journal of Educational Technology in Higher Education*, 13, 20. https://www.doi.org/10.1186/s41239-016-0019-9

Roßnagel, C. S., Fitzallen, N. & Lo Baido, K. (2021). Constructive alignment and the learning xperience: relationships with student motivation and perceived learning demands. *Higher Education Research & Development, 40*, 838–851. https://doi.org/1.0.1080/07294360. 2020.1787956

Rupp, A. A. & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66,* 63–84. https://doi.org/10.1177/0013164404273942

Samejima, F. (2016). Graded response model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Models* (S. 95–108). Boca Raton, FL: Chapman & Hall/CRC.

Schaper, N. & Hilkenmeier, R. (2013). *Umsetzungshilfen für kompetenzorientiertes Prüfen.* Zugriff am 02.03.2021. Verfügbar unter https://www.hrk-nexus.de/fileadmin/redaktion/hrk-nexus/07-Downloads/07-03-Material/zusatzgutachten.pdf

Schmees, M. & Horn, J. (2014). *E-Assessments an Hochschulen: Ein Überblick. Szenarien. Praxis. E-Klausur-Recht*. Münster: Waxmann.

Schulz, A. (2016). E-Examinations: Zur Computerisierung des Prüfungswesens deutscher Hochschulen. *Forschung & Lehre*, *3*, 208–209.

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 429–438). Boston: Elsevier Academic. https://doi.org/10.1016/b0-12-369398-5/004448

Spoden, C. & Fink, A. (2021). Sicherheitsaspekte bei der Konstruktion und Administration von Hochschulklausuren. In C. Spoden & A. Frey (Hrsg.), *Psychometrisch fundierte E-Klausuren für die Hochschule* (S. 49–54) Lengerich: Pabst Science Publishers.

Spoden, C. & Frey, A. (Hrsg.) (2021). *Psychometrisch fundierte E-Klausuren für die Hochschule*. Lengerich: Pabst Publishers.

Stocking, M. L. (1988). Scale drift in online calibration. *ETS Research Report Series, 1988*(1), 1–122. https://doi.org/10.1002/j.2330-8516.1988.tb00284.x

Stocking, M. L. (1997). Revising Item Responses in Computerized Adaptive Tests: A Comparison of Three Models. *Applied Psychological Measurement, 21*, 129–142. https://doi.org/10.1177/01466216970212003

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210. https://doi.org/10.1177/014662168300700208

St-Onge, C., Quellett, K., Lakhal, S., Dubé, T. & Marceau, M. (2021). COVID-19 as the tipping point for integrating e-assessment in higher education practices. *British Journal of Educational Technology, 53*, 349–366. https://doi.org/10.1111/bjet.13169

Terzis, V. & Economides, A. A. (2011). The acceptance and use of computer based assessment. *Computers & Education, 56,* 1032–1044. https://doi.org/10.1016/j.compedu.2010.11.017

Terzis, V., Moridis, C. N. & Economides, A. A. (2012). How student's personality traits affect Computer Based Assessment Acceptance: Integrating BFI with CBAAM. *Computers in Human Behavior, 28,* 1985–1996. https://doi.org/10.1016/j.chb.2012.05.019

Tonidandel, S., Quinones, M. A. & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test-takers' reactions*. Journal of Applied Psychology, 87,* 320–332. http://doi.org/10.1037//0021-9010.87.2.320

van der Linden, W. J. (2016a). Unidimensional logistic response models. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume one: Models* (pp. 13–30). Boca Raton, FL: Chapman & Hall/CRC.

van der Linden, W. J. (Ed.) (2016b). *Handbook of item response theory, Volume one: Models*. Boca Raton, FL: Chapman & Hall/CRC. https://doi.org/10.1201/9781315374512

van der Linden, W. J. (2018). Adaptive testing. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume three: Applications* (pp. 197–228). Boca Raton, FL: Chapman & Hall/CRC.

van der Linden, W. J. (2021). Review of the shadow-test approach to adaptive testing. *Behaviormetrika*, *48*. https://doi.org/10.1007/s41237-021-00150-y

van der Linden, W. J. & Choi, S. W. (2019). Improving item-exposure control in adaptive testing. *Journal of Educational Measurement*, *57*, 405–422. https://doi.org/10.1111/jedm.12254

van der Linden, W. J. & Glas, C. A. W. (Eds.) (2010). *Elements of adaptive testing*. New York, NY: Springer. https://doi.org/10.1007/978-0-387-85461-8

van der Linden, W. J. & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In van der Linden, W. J. & Glas, C. A. W. (Eds.), *Elements of adaptive testing* (pp. 3–30). New York, NY: Springer. https://doi.org/10.1007/978-0-387-85461-8_1

van der Linden, W. J. & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*, 259–270. https://doi.org/10.1177%2F014662 16980223006

van der Linden, W. J. & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29, 273–291.* https://doi.org/10.3102%2F10769986029003273

van der Linden, W. J. & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32*, 398–418. https://doi.org/10.3102%2F1076998606298044

Venkatesh, V., Morris, M., Davis, G. & Davis, F. (2003). User acceptance of information technology: towards a unified view. *MIS Quarterly, 27*, 479–501. https://doi.org/10.2307/30036540

Verschoor, A., Berger S., Moser U. & Kleintjes F. (2019). On-the-Fly Calibration in Computerized Adaptive Testing. In B. Veldkamp & C. Sluijter (Eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement: Methodology of Educational Measurement and Assessment* (pp. 307–323). Cham: Springer. https://doi.org/10.1007/978-3-030-18480-3_16

Wang, S., Fellouris, G. & Chang, H. H. (2017). Computerized adaptive testing that allows for response revision: Design and asymptotic theory. *Statistica Sinica, 27,* 1987–2010. https://doi.org/10.5705/ss.202015.0304

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika*, *54*, 427–450. https://doi.org/10.1007/BF02294627

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing, 2,* 1–17. http://doi.org/10.7333/1401-0201001

Wise, S. L. (2019). Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety & cheating. *Education Inquiry, 10,* 21–23. https://doi.org/10.1080/20004508.2018.1490127

Wise, S. L. (2020). An intelligent CAT that can deal with disengaged test taking. In H. Jiao & R. W. Lissitz (Eds.), *Application of artificial intelligence to assessment* (pp. 161–174). Charlotte, NC: Information Age Publishing.

Wise, S. L., Bhola, D. & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25*(2), 21–30. https://doi.org/10.1111/j.1745-3992.2006.00054.x

Wise, S. L. & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education, 30*, 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wu, Q., Debeer, D., Buchholz, J., Hartig, J. & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-scale Assessments in Education, 7(1)*, 5. https://doi.org/10.1186/s40536-019-0073-6

Yan, D., Rupp, A. A. & Foltz, P. W. (Eds.) (2020). *Handbook of automated scoring: Theory into practice*. Boca Raton, FL: Chapman & Hall/CRC.

Zheng, M. & Bender, D. (2019). Evaluating outcomes of computer-based classroom testing: Student acceptance and impact on learning and exam performance. *Medical Teacher, 41*(1), 75–82. https://doi.org/10.1080/0142159X.2018.1441984

Anhang

Anhang A: Beitrag 1 – A continuous calibration strategy for computerized adaptive testing

Zitation: Fink, A., Born, S., Spoden, C. & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling,* 60, 327–346. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_3-2018_327-346.pdf

Psychological Test and Assessment Modeling, Volume 60, 2018 (3), 327-346

A continuous calibration strategy for computerized adaptive testing

Aron Fink[1], Sebastian Born[2], Christian Spoden[2,3] & Andreas Frey[2,4]

Abstract

This paper presents a new continuous calibration strategy for using computerized adaptive testing in application areas where it is not feasible to conduct a separate calibration study and/or to construct the complete item pool before the operational phase of the test. This method enables a step-by-step build-up of the item pool across several test cycles. A combination of equating and linking is used to maintaining the scale across these cycles. A simulation study was carried out to investigate the performance of the strategy regarding the precision of the ability estimates. The simulation study is based on a full factorial design with the factors IRT model, sample size and number of new uncalibrated items added to the item pool per test cycle. Precision of the ability estimates increased over the test cycles in all conditions. For the 2PL model, a better performance was reached when using a lower number of new uncalibrated items. The results support the application of the new method especially in small sample sizes.

Keywords: computerized adaptive testing, item response theory, online calibration, item banks, test design

[1]*Correspondence concerning this article should be addressed to:* Aron Fink, Institute of Educational Science, Department of Research Methods in Education, Friedrich Schiller University Jena, Am Planetarium 4, 07743 Jena, Germany, email: aron.fink@uni-jena.de

[2]Friedrich Schiller University Jena, Germany

[3]Now at the German Institute for Adult Education – Leibniz Centre for Lifelong Learning, Bonn, Germany

[4]Centre for Educational Measurement (CEMO) at the University of Oslo, Norway

Computerized adaptive testing (CAT) is a testing mode in which the selection of the item to be presented next to the test taker depends upon the responses given to previously administered items (Frey, 2012). Extensive prior research shows that CAT typically yields more precise ability estimates and/or a shorter test length compared to traditional nonadaptive testing (e.g., Segall, 2005). In addition, CAT has a great potential to overcome some limitations of traditional nonadaptive tests, which could explain its growing popularity in many different fields, such as psychological and educational testing, large-scale assessments, admission testing, health outcome assessments and others (for a list of operational CAT programs see http://www.iacat.org/content/operational-cat-programs). In particular, CAT can be designed so that test takers are measured with a comparable level of precision across the complete ability range (Frey & Ehmke, 2007). In contrast, traditional nonadaptive tests typically provide the highest precision for test takers of medium ability, while the precision decreases for test takers with test scores in the extremes (Dolan & Burling, 2012).

An essential building block in the development of a computerized adaptive test is a calibrated item pool (e.g., He & Reckase, 2014; Thompson & Weiss, 2011). Traditionally, one single calibration study is carried out, in which a large number of test takers respond to a large number of candidate items. Based on the responses gathered in the calibration study, item parameters are estimated by means of item response theory (IRT; e.g., van der Linden, 2016) methods. In the subsequent operational phase of a computerized adaptive test, the estimated item parameters are considered to be known and are used for item selection and ability estimation. Therefore, the quality of a computerized adaptive test depends, inter alia, on the precision of the item parameter estimates, which will, in turn, be determined mainly by the number of responses per item (i.e., the calibration sample size and the test length). For the frequently used two-parameter logistic (2PL) test model, for example, a minimum of 500 responses per item is recommended (de Ayala, 2009). Unfortunately, in various potential application areas of CAT, constructing large numbers of items prior to the initial use of the test (see Spoden, Frey and Bernhardt, in press, for a special case where using already field tested items was an option) and/or carrying out a calibration study with a large sample is not feasible, due to a lack of resources available for test development. Correspondingly, for applications such as written exams, psychological tests used in personnel selection, for clinical diagnosis or in research, CAT is typically not used, even though it would be advantageous here as well. In addition, test developers often access test takers from low-stakes samples for the calibration study, even if the intended population of test takers comes from another context. This procedure comes with a limitation. There is evidence that motivational differences exist between the low-stakes calibration samples and the intended high-stakes population of test takers (e.g., Sundre & Kitsantas, 2004; Wise & DeMars, 2005). These motivational differences may result in biased item parameter estimations in the calibration phase, which can jeopardize the interpretation of the test scores.

A possible solution to overcome these problems is the method of online calibration (e.g., Stocking, 1988; Kingsbury, 2009; Chen, 2017). Originally, online calibration methods were developed to pretest and calibrate new items on the fly during the administration of a computerized adaptive test. Most of the existing online calibration methods are, however,

based on the assumption that an operational computerized adaptive test already exists and online calibration is only used to add new items to the item pool. Until now, the paper of Makransky and Glas (2010) is the only available study that uses online calibration designs to calibrate an item pool without having any information about the item parameters at the beginning of the operational CAT phase. Makransky and Glas (2010) presented a continuous updating calibration strategy, in which item parameters are re-estimated after each item administration using all previous responses. Therefore, the precision of the item and thus the precision of the ability estimates are continuously improved.

This strategy can be problematic for two major reasons. First, especially for high-stakes testing situations, re-estimating item parameters after each item administration leads to legal problems because ability measures are estimated from different item parameters and cannot be compared on the same established scale. This strategy is sound from a psychometric point of view but would be hard to defend in the court in case of a lawsuit. Second, in some cases the scales have to be linked across test cycles in order to make it possible to directly compare the test results across test cycles.

Against this background, a calibration strategy is necessary that allows for a simultaneous item and ability estimation and ensures a fair comparison of measures across test takers from different test cycles. The aim of the present study is thus to propose and examine a new strategy to calibrate items continuously across several test cycles without a separate calibration study, increasing the item pool across test cycles and maintaining the scale in each of these cycles.

Model

The present study was carried out in the framework of unidimensional logistic response models for dichotomous items. This family of models has a long tradition and is widely used in the field of educational and psychological testing (van der Linden, 2016). One model used in this study is the two-parameter logistic (2PL) model. This model defines the probability of a correct response $u_{ij} = 1$ of person $j = 1, \ldots, N$ with the latent ability level θ_j to an item i as:

$$P\left(u_{ij} = 1 \middle| \theta_j, a_i, b_i\right) = \frac{\exp\left[\, a_i\left(\theta_j - b_i\right)\right]}{1 + \exp\left[\, a_i\left(\theta_j - b_i\right)\right]}, \tag{1}$$

where a_i is the discrimination parameter and b_i is the difficulty parameter of item i. Equation (1) reduces to the one-parameter logistic (1PL) model when the discrimination parameter a_i is set to a constant for all items (van der Linden, 2016). It should be noted that especially in high-stakes testing situations with closed response formats, the assumption of nonguessing, and thus a pseudoguessing parameter equal to zero is rather strict. However, the 2PL and the 1PL model are viable alternatives to more complex models because estimation of additional item parameters (e.g., pseudoguessing parameter) can be troublesome in small samples and tends to be less stable across different assessments. Unstable item parameter estimates complicate the process of linking across different test cycles,

which makes it difficult to compare ability measures across these test cycles. Thus, the 2PL and the 1PL model are used in this study.

Continuous calibration strategy

The proposed continuous calibration strategy is divided into two phases, the *initial phase* and the *continuous phase*. The initial phase describes the first test cycle. During the initial phase, the same set of items is administered to every test taker. In fact, there is no difference between the initial phase of this procedure and a traditional nonadaptive test. Based on the item responses, item parameters and person parameters are estimated.

All test cycles following the initial phase are subsumed under the continuous phase. The major difference between the two phases is the type of item administration, which is nonadaptive in the initial phase and partly adaptive in the continuous phase. Tests in the continuous phase consist of three item clusters named *adaptive cluster*, *calibration cluster*, and *linking cluster*. The adaptive cluster contains items that are administered adaptively using the item parameter estimates from the previous test cycle. In the calibration cluster, items with unknown item parameters are included to enlarge the item pool. These new items are administered to every test taker of the current test cycle. Finally, the linking cluster comprises items that have already been administered in previous test cycles and have the most accurate item parameter estimates, compared to the other items in the item pool. These items are used to link consecutive test cycles with each other, using a common-item nonequivalent group design (Kolen & Brennan, 2014). The linking cluster makes it possible to report the results obtained in the various test cycles on the same scale and thereby allows direct comparisons of test results across these cycles. As a rule of thumb, at least 20 % of the test items should be used for linking, and each cluster of link items should be a good representation of the total test, both in content and statistical characteristics (Kolen & Brennan, 2014). To establish a stable linking between consecutive test cycles, it is necessary that the parameters of the link items are invariant across these cycles. Therefore, checking the link items for item parameter drift is necessary (IPD; Goldstein, 1983; Bock, Muraki, & Pfeiffenberger, 1988; Wells, Subkoviak, & Serlin, 2002). IPD occurs when the invariance assumption of the item parameters no longer holds across two or more test administrations. Link items showing significant differences in their item parameter estimates between two test cycles have to be identified and excluded from the linking procedure.

Each test cycle in the continuous phase can be divided into seven steps that are illustrated in Figure 1. First, items for the linking cluster of the current test cycle are selected (step 1). Subsequently, a test that is composed of an adaptive, a calibration and a linking cluster is administered (step 2). To check link items for IPD, item parameters are initially estimated based on the responses of the current test cycle (step 3). Since the examinees of the previous and current test cycles are not considered to be equivalent, parameter estimates for the two estimations are not on the same scale and therefore not directly comparable. Thus, a scale transformation has to be conducted. The link item parameters obtained from step 3 are brought onto the same scale as their parameters from the preceding test cycle by

means of scale transformation methods (step 4; e.g., Kolen & Brennan, 2014). Thereafter, the set of link item parameter estimates is checked for IPD (step 5). Link items showing significant drift are excluded from linking and estimated freely in step 6a/6b. The test for IPD is accomplished iteratively (iterative purification method; Wells, Hambleton, Kirkpatrick, & Meng, 2014). After excluding link items showing IPD, the scale transformation of the remaining link items (step 4) has to be repeated again. This could lead to different results of the test for IPD, and, if necessary, to the exclusion of additional link items. The iterative process stops whenever there is no longer a link item showing significant IPD. Subsequently, item parameters are estimated using a fixed item parameter calibration based on the responses gained from all previous test cycles, whereby the item parameters of the remaining link items from step 5 are fixed at their estimates from the preceding test cycle (step 6b). Whenever a complete breakdown of the link occurs (i.e., there are not enough link items left to establish a stable linking), all item parameters are estimated freely based on the responses gained from all previous test cycles (step 6a). This type of estimation is often referred to as concurrent calibration (Wingersky & Lord, 1984). In this case, the results are not on the same scale as the results of the previous test cycles anymore, and, therefore, the linking procedure has to be started anew. Note that at least two link items need to be stable to keep the location and variation of the scale comparable across test cycles. However, it should be noted that the fewer the items used for the linking, the more prone to sampling errors the linking procedure is (Wingersky & Lord, 1984). Therefore, test developers should wisely choose the minimum number of items needed for fixed item parameter calibration. Finally, the ability parameters are estimated based on the item parameters from step 6a/6b (step 7).

To sum up, by applying the continuous calibration strategy, the item pool size increases by adding new items to the item pool, the item parameters are updated continuously while maintaining the reporting scale by means of the linking procedure, and the test gains precision and adaptivity across test cycles. In addition, differentiating between the three item clusters adds a lot of flexibility to the algorithm. If increasing the item pool size is not a primary interest anymore, the calibration cluster becomes an additional adaptive cluster and, therefore, adaptivity is maximized.

The continuous calibration strategy in its basic form combines different psychometric approaches and therefore has many tuning parameters (e.g., sample size per test cycle, underlying IRT model, test length, number of items per cluster, link item selection method, test for IPD). To find the best configuration of the proposed strategy and to give practical recommendations, these tuning parameters should be examined prior its operational use.

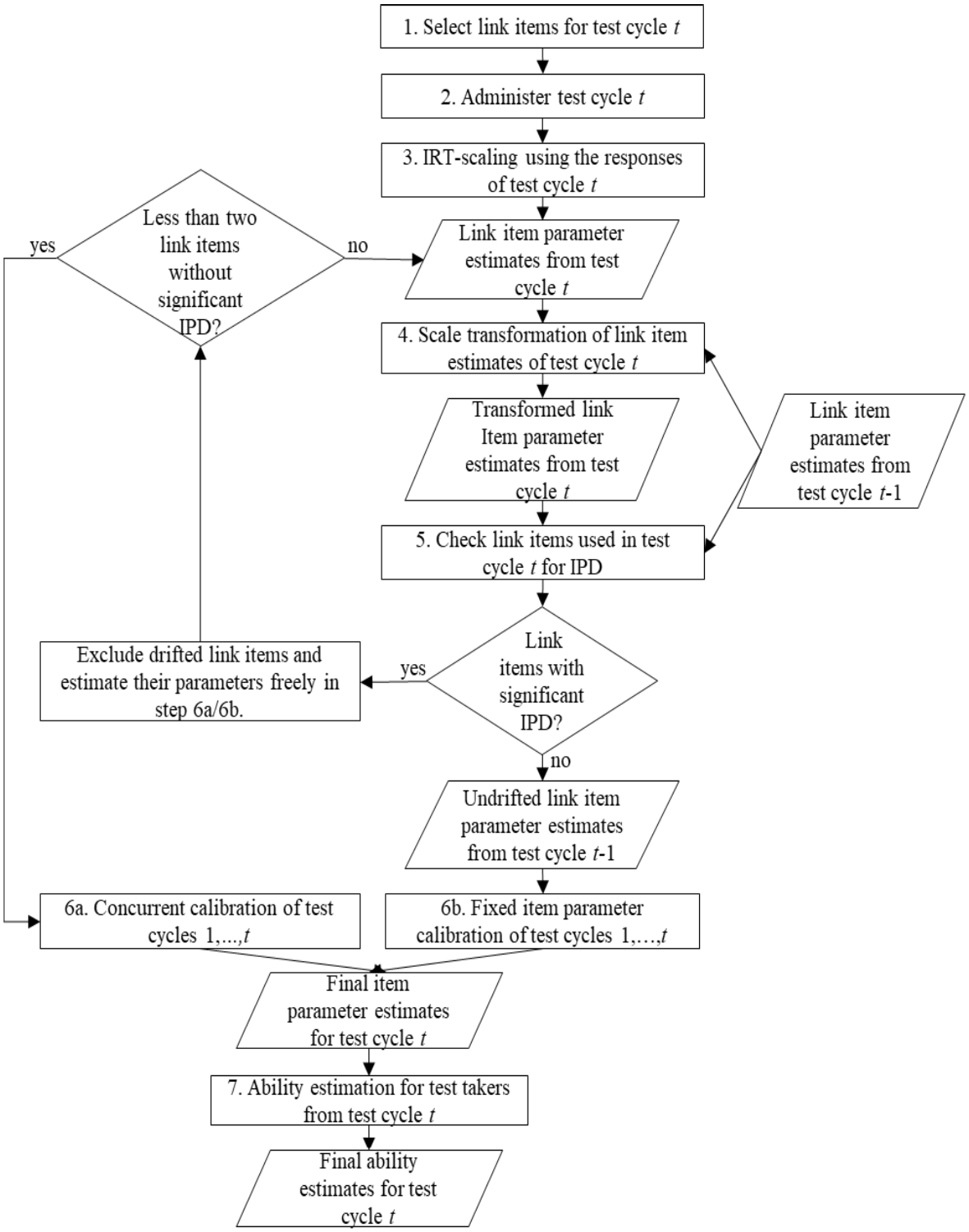


Figure 1:
Flowchart of the continuous phase of the proposed continuous calibration strategy.

Research questions

The main aim of the following simulation study is to investigate the performance of the strategy regarding the quality of the obtained ability estimates in general and to give practical recommendations for the configuration of the algorithm. As there are numerous different possible configurations of the continuous calibration strategy, in this study only some of the basic tuning parameters are varied while others are kept constant. The first one is the sample size per test cycle. As already mentioned above, the number of test takers has a strong influence on the precision of the item parameter estimates. However, there are different recommendations concerning the required sample size for item calibration (e.g., "a few hundred"; de Ayala, 2009, p. 43). As the continuous calibration strategy is supposed to be applicable in areas where only small sample sizes are available for calibration purposes, the effect of different, rather small sample sizes on the performance of the continuous calibration is of interest. Therefore, the first research question is as follows:

1. What is the effect of using different sample sizes per test cycle in the continuous calibration strategy on the precision of the ability estimates at a different number of test cycles?

The second tuning parameter to be varied in this study is the number of items in the calibration cluster (which in turn means the number of test cycles necessary to obtain an initial estimation for each item in an item pool of a prespecified size), labeled *calibration speed.* This tuning parameter is supposed to strongly influence the performance of the continuous calibration strategy. Given a predefined test length, a higher number of items in the calibration cluster limits the number of items in the adaptive cluster. In addition, this leads to a faster increase in the number of items the selection algorithm can choose from. Thus, the level of calibration speed determines the level of adaptivity in each test cycle. In contrast, enlarging the item pool slowly leads to a more uniform item administration, which leads to a higher number of responses per item, and this in turn leads to more precise parameter estimates. The effect of using different levels of calibration speed is thus not easily predictable. Consequently, the second research question is:

2. What is the effect of using different levels of calibration speed in the continuous calibration strategy on the precision of the ability estimates at a different number of test cycles in the calibration process?

The last tuning parameter to be varied in this study is the underlying IRT model. As already mentioned above, in this study the 1PL and the 2PL model are used. Since the discrimination parameters in the 2PL model tend to be less stable than the difficulty parameters, a larger number of responses per item is recommended for calibration compared to the 1PL model. Bearing in mind that the number of responses per item in this study depends on the sample size and the calibration speed, it can be expected that there are differences in the results regarding research questions 1 and 2 when using different IRT models. Therefore, the last research question is:

3. To what extent does the chosen IRT model in the continuous calibration strategy affect the results for research questions 1 and 2?

Method

A study design with three independent variables was used. With the first independent variable, *Sample Size*, the number of test takers per test cycle ($N = 50$, $N = 100$, $N = 300$) was varied. The second independent variable, *Calibration Speed* ($t = 3$, $t = 5$, $t = 9$), compares the number of test cycles t necessary to obtain an initial estimation for each item in the item pool. Please note that concerning the calibration speed, the size of the item pool to be calibrated was 130 items, the test length for one test cycle was set to 50 items, and the number of items in the linking cluster for one test cycle was set to 10 items. For $t = 9$, every test in the continuous phase comprised 30 items in the adaptive cluster and 10 items in the calibration cluster. Since the item pool increased by 10 items in each test cycle, the complete item pool was calibrated after nine test cycles. This condition represents the slowest calibration procedure used in this study and determines the number of test cycles the continuous calibration was running in each condition. Thus, in order to keep the overall number of responses constant across conditions, the number of test cycles was set to nine for each condition. For $t = 5$, the tests in the continuous phase contained 20 items in the adaptive cluster and 20 items in the calibration cluster. Thus, after five test cycles every item had an initial estimation. From test cycle six to nine the calibration cluster became an additional adaptive cluster. For $t = 3$, the tests in the continuous phase comprised 10 items in the linking cluster and 40 items in the calibration cluster. No items were administered adaptively until test cycle four. From test cycle four to nine, the calibration cluster became the adaptive cluster. This condition represents the fastest possible calibration procedure, given the test specifications. The third independent variable, *IRT Model* (1PL, 2PL), represents the underlying IRT model used for calibration.

The fully crossed design had 2 x 3 x 3 = 18 conditions. For each of the conditions, 200 replications were analyzed with regard to the measurement precision as subsequently defined.

The simulation was carried out in R (R Core Team, 2017) using the "mirtCAT" package (Chalmers, 2016) for simulating the adaptive tests and the "mirt" package (Chalmers, 2012) for item and person parameter estimation. These functions were called from R-code that was written anew to carry out the continuous calibration strategy.

Simulation procedure

For each replication, the ability parameters were randomly drawn from a standard normal distribution, $\theta \sim N(0,1)$. The b_i parameters for each replication were drawn from a truncated normal distribution $b_i \sim N(0, 1.5)$, $b_i \in (-4.5, 4.5)$. This kind of distribution was chosen, because it can be assumed that under real conditions item parameters are usually located within this interval. The a_i parameters were drawn from a lognormal distribution, $a_i \sim lognormal(0, 0.25)$.

Items in the calibration cluster and the linking cluster were administered sequentially; items in the adaptive cluster were selected using the maximum information criterion (Lord, 1980) based on the item parameter estimates obtained in the preceding test cycle.

After each test cycle, the items were calibrated under either the 1PL or the 2PL model using marginal maximum likelihood (MML, Bock & Aitkin, 1981) estimation and subsequently, person parameters were estimated using weighted maximum likelihood estimation (WLE; Warm, 1989). This method was preferred over maximum likelihood estimation, because it is less biased and provides ability estimates for test takers with invariant response patterns.

Selection of link items

For the selection of link items, previously administered items were categorized based on their estimated difficulty parameters. The interval limits of the categories were determined as quantiles of the item difficulty distribution:

Category 1 (very low difficulty): $b_i \in (b_{min}, b_{.1}]$;

Category 2 (low difficulty): $b_i \in (b_{.1}, b_{.3}]$;

Category 3 (medium difficulty): $b_i \in (b_{.3}, b_{.7}]$;

Category 4 (high difficulty): $b_i \in (b_{.7}, b_{.9}]$;

Category 5 (very high difficulty): $b_i \in (b_{.9}, b_{max})$.

Within each of these five categories, items with the lowest standard error (*SE*) of the difficulty parameter estimate were selected as link items. One item from category 1, two items from category 2, four items from category 3, two items from category 4 and one item from category 5 were selected to serve as potential link items. This procedure ensured that the distribution of the link items resembled the distribution of the complete item pool and, therefore, the linking cluster was a good representation of the whole test in terms of statistical characteristics. Before using these items for the linking, they were tested for IPD (s. Figure 1). For this purpose, item parameter estimates obtained from step 3 of the continuous calibration strategy were brought up to the same scale as their parameter estimates from the preceding test cycle. There are several popular transformation methods that can be used: mean/mean (Loyd & Hoover, 1980), mean/sigma (Marco, 1977), item characteristic curves (Haebara, 1980), and test characteristic curves method (Stocking & Lord, 1983). In this study, the mean/mean method was chosen due to the simple and user-friendly implementation of the method in comparison to the characteristic curve methods. In addition, the mean/mean method was preferred over the mean/sigma method, because means are typically more stable than standard deviations, as noted by Baker and Al-Karni (1991), and mean/sigma ignores the information from the a-parameters for the 2PL model. It should be noted that if the test is suspected to have very atypical item parameter estimates, the use of the moment methods (mean/mean, mean/sigma) may be questionable since they are highly sensitive to item parameter outliers. For practitioners, Hanson and Béguin (2002) suggested that it would be beneficial to apply multiple linking procedures and compare the scaling results. However, for the purpose of this study, it is sufficient to apply only one easily implementable transformation method. After scale transformation, it was examined whether the item parameter estimates of a link item from the preceding test cycle falls into the 95 % confidence interval around the same item's parameter estimates from

the current test cycle. If the difficulty parameter estimate, the discrimination parameter estimate or both fell outside the confidence interval, the assumption of item parameter invariance was rejected for this item and, consequently, this item was not used for fixed item parameter calibration. In addition to this relatively simple and easily implementable method, there are also more complex methods for detecting IPD (e.g., mixed distribution IRT models; Park, Lee, & Xing, 2016), which, however, set stronger demands on the data, especially regarding the required sample size. The test for IPD was done iteratively. Drifted link items were excluded from linking and estimated freely without imposing any constraints.

Evaluation criteria

The mean squared error (MSE) of the ability estimates was used to evaluate the precision of the ability estimation after each test cycle for each of the conditions. The global MSE was computed as the average squared difference between the ability estimates $\hat{\theta}_j$ and the true ability θ_j for the N test takers in each test cycle and was averaged across the $rep = 200$ replications:

$$MSE = \frac{1}{rep * N} \sum_{r=1}^{rep} \sum_{j=1}^{N} (\hat{\theta}_j - \theta_j)^2. \quad (2)$$

Low MSE indicates high measurement precision. In addition to the global MSE (Equation 2), the precision of the ability estimates given a particular ability level was also investigated. This conditional MSE was calculated for a specific range of ability measures after each test cycle. The seven ability ranges used for this procedure were $\theta_j \in (-Inf, -2]$, $\theta_j \in (-2, -1]$, $\theta_j \in (-1, 0.25]$, $\theta_j \in (-0.25, 0.25]$, $\theta_j \in (0.25, 1]$, $\theta_j \in (1, 2]$, and $\theta_j \in (2, Inf)$.

Furthermore, a lower (L) and an upper baseline (U) were simulated to compare the continuous calibration design to additional criteria. For the lower MSE baseline, $\hat{\theta}_j$ was estimated after each test cycle using the real item parameters to simulate the hypothetically best possible estimation without errors stemming from the estimation of the item parameters. The precision of $\hat{\theta}_j$ stemming from item parameters that were fixed after their first estimation and thus not updated continuously was set as the upper MSE baseline.

Results

Global precision

To answer the research questions, in a first step the global MSE after each test cycle was analyzed. Figure 2 shows the MSE across test cycles under the 1PL for all sample sizes and calibration speeds. As seen, the MSE decreased over the test cycles in all conditions. Only during the first three test cycles for $t = 3$ did the MSE remain constant. This is caused

by the fact that in this condition the first three test cycles are linked nonadaptive tests. For $t = 3$ and $t = 5$ the MSE dropped at the $(t + 1)^{th}$ test cycle, which could be explained by the increasing adaptivity stemming from the conversion of the calibration cluster into an additional adaptive cluster. Up to the $(t + 1)^{th}$ test cycle the respective slower calibration procedure showed a lower MSE, especially for $N = 50$. For the ninth test cycle, the different levels of calibration speed seem to perform similarly for each sample size with a little disadvantage of the $t = 9$ condition, which could be explained by the fact that there is still a calibration cluster in the ninth test cycle and therefore less adaptivity in this condition only.

Over the course of the test cycles, the MSE converged to the lower baseline and moved away from the upper baseline in all conditions. However, for $N = 300$ there were nearly no differences between the MSE stemming from the continuous calibration and the upper and the lower baseline. For this sample size, using the initial estimation of the item parameters (without updating the parameters) works approximately as well as using the real item parameters. The increasing precision is apparently based only on the growth of the item pool.

Figure 3 shows how the MSE evolved over the test cycles under the 2PL model. The results resemble the results of the 1PL, with the exception that the lowest MSE was reached for $t = 9$ over the course of the nine test cycles. In addition, for $t = 3$, the MSE increased from test cycle one to test cycle two even for $N = 300$, which could be interpreted as an effect of the linking procedure.

Conditional precision

In a second step, the conditional precision of the ability estimates for specific ranges of the ability scale was investigated. Figure 4 illustrates the results for the 1PL. For the sake of clarity of presentation, the results are only presented for the first, the third, the fifth and the ninth test cycles. As seen, in every condition and in all test cycles in the calibration process, the precision of ability estimates was highest for medium level ability scores. Over the course of the test cycles, the difference in measurement precision between test takers with extreme ability scores and test takers with medium ability scores decreased. For t = 3, the results of the first and the third test cycles, similar to the results of the fifth and the ninth test cycles, were fairly equal. Updating the item parameters in this condition had only very small impact on the conditional precision.

Figure 5 illustrates the results under the 2PL model. These were very similar to those obtained under the 1PL model, but for $t = 3$, similar to global precision, the conditional precision decreased from test cycle one to three. Between $t = 5$ and $t = 9$ there were nearly no differences in the conditional precision over the course of the test cycles.

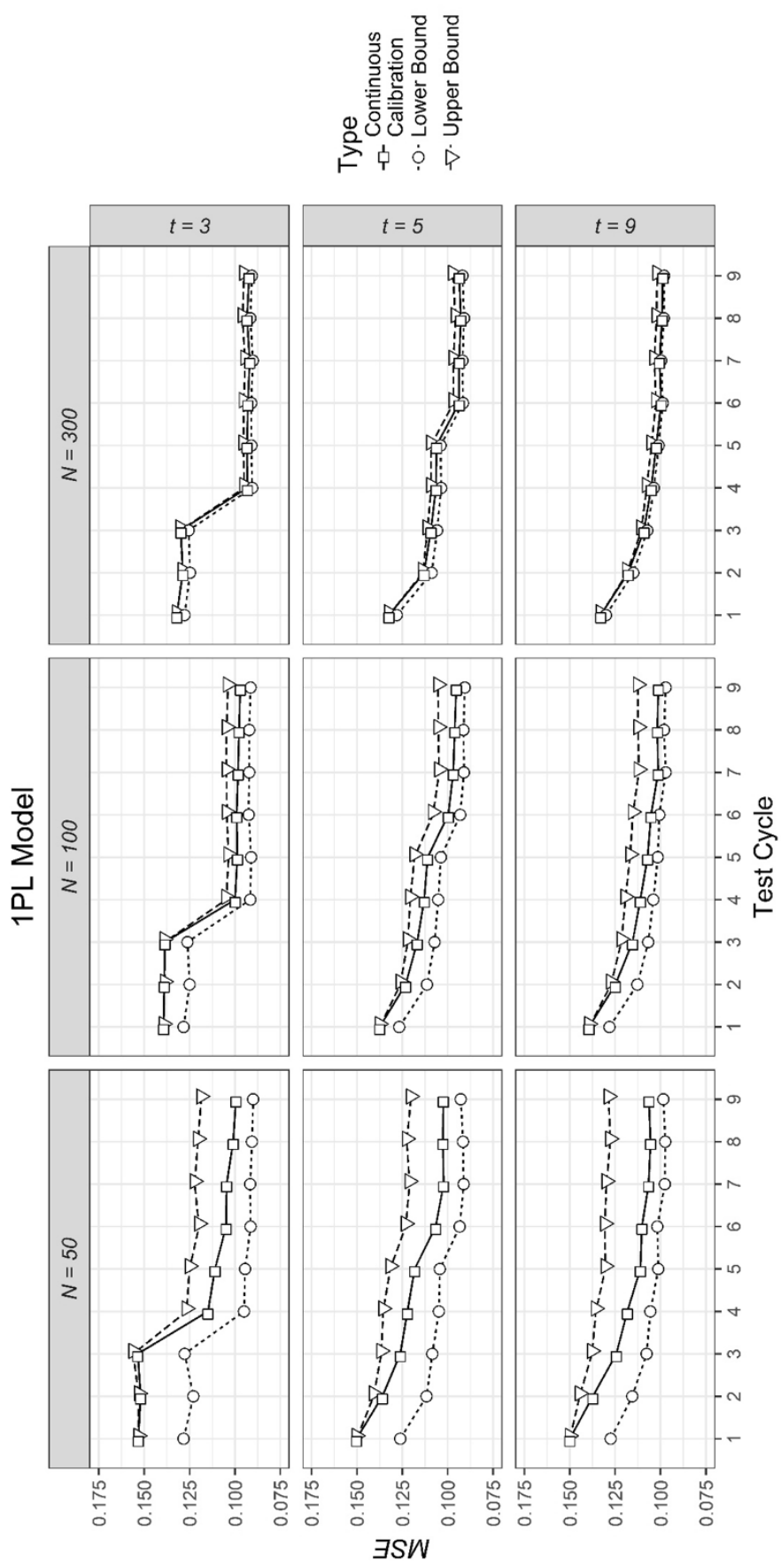


Figure 2:
Global mean squared error (*MSE*) for each test cycle (*t*) in the continuous calibration strategy for different sample sizes and different levels of calibration speed under the 1PL model.

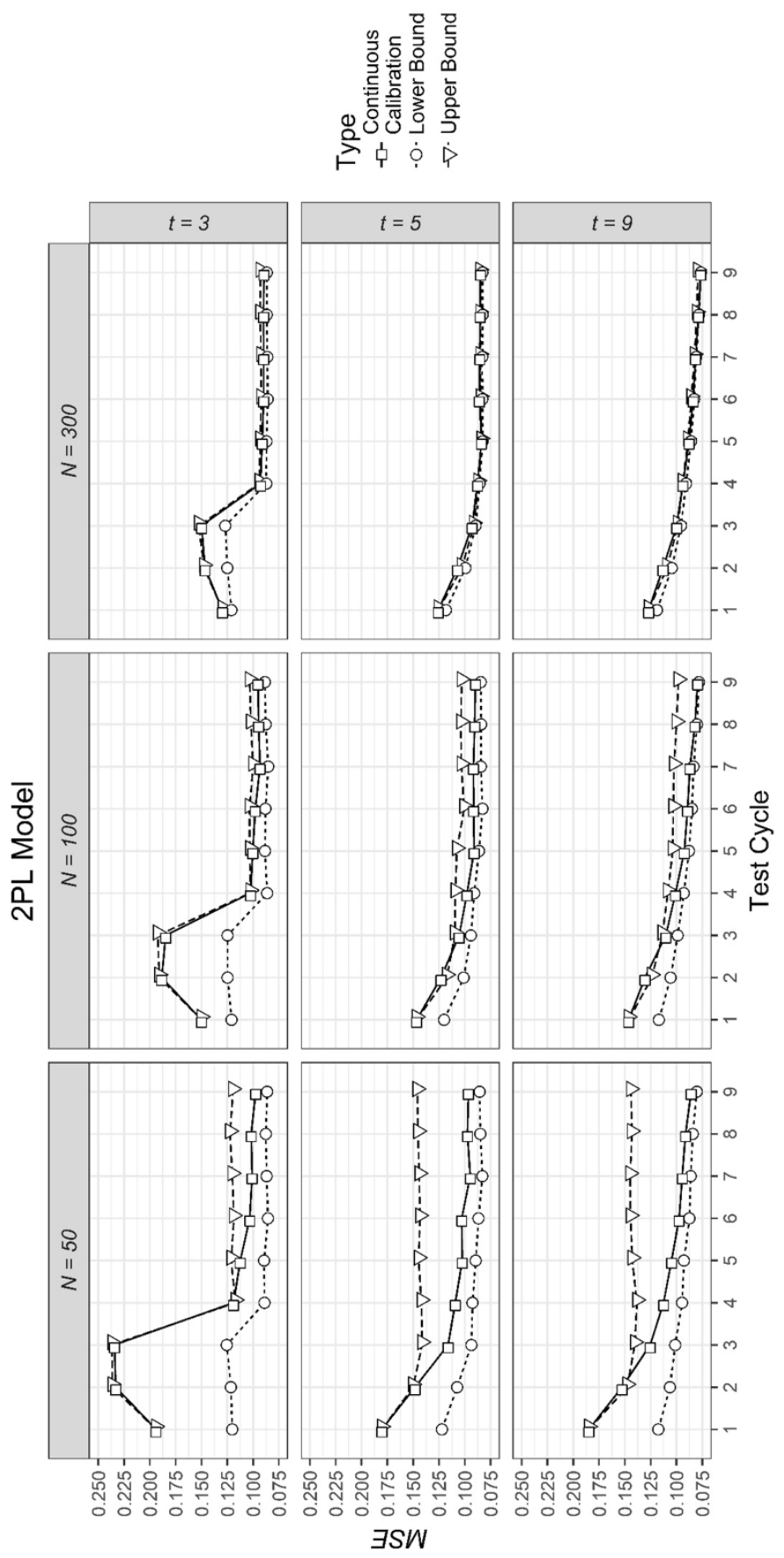


Figure 3:
Global mean squared error (*MSE*) for each test cycle (*t*) in the continuous calibration strategy for different sample sizes and different levels of calibration speed under the 2PL model.

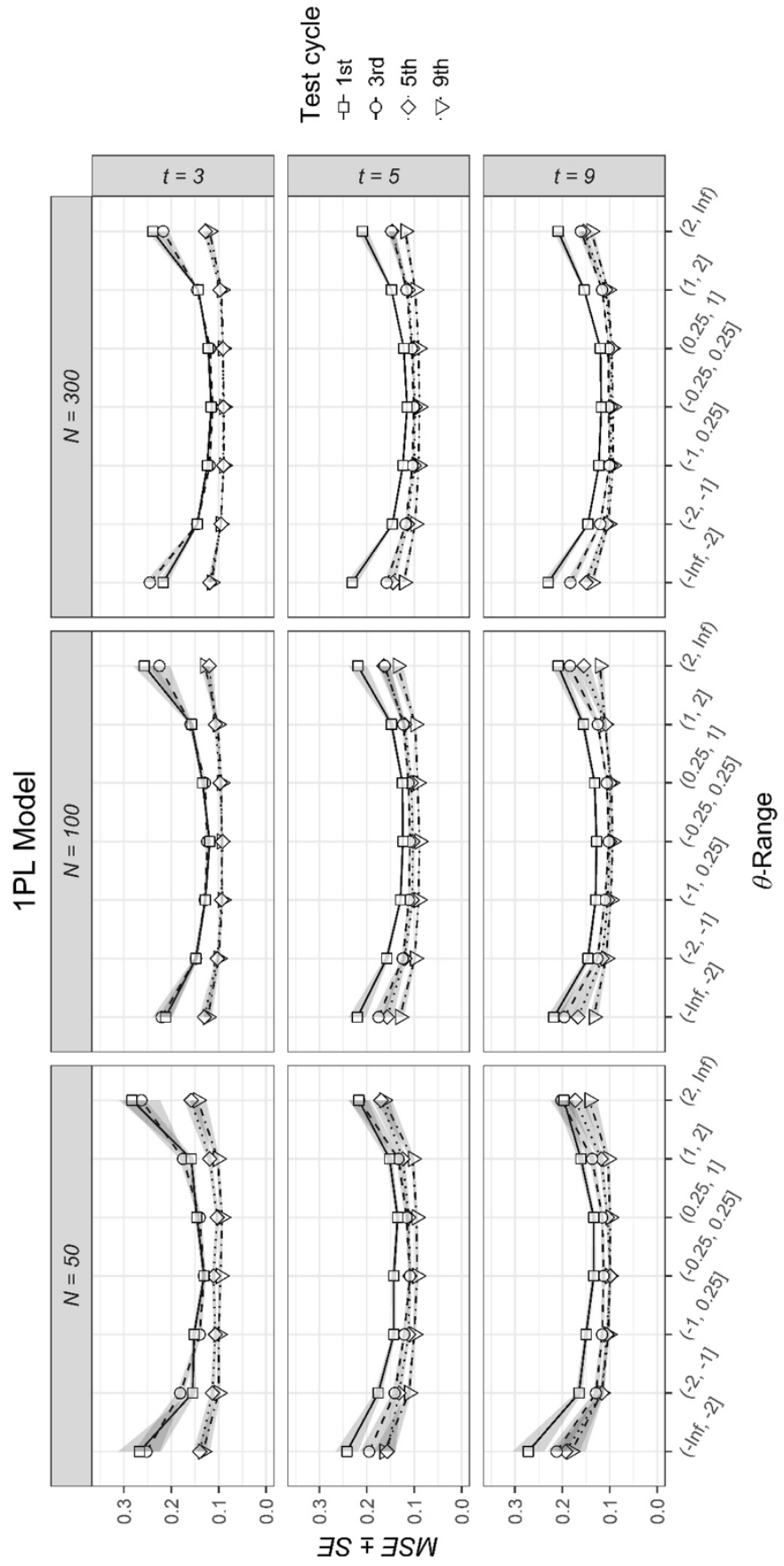


Figure 4:
Conditional mean squared error (*MSE*) at specific areas on the θ-continuum after different numbers of test cycles in the continuous calibration strategy for different sample sizes and different levels of calibration speed under the 1PL model. The areas shaded gray represent the standard error around the calculated MSE obtained from the variability of all simulees in a specific θ-range.

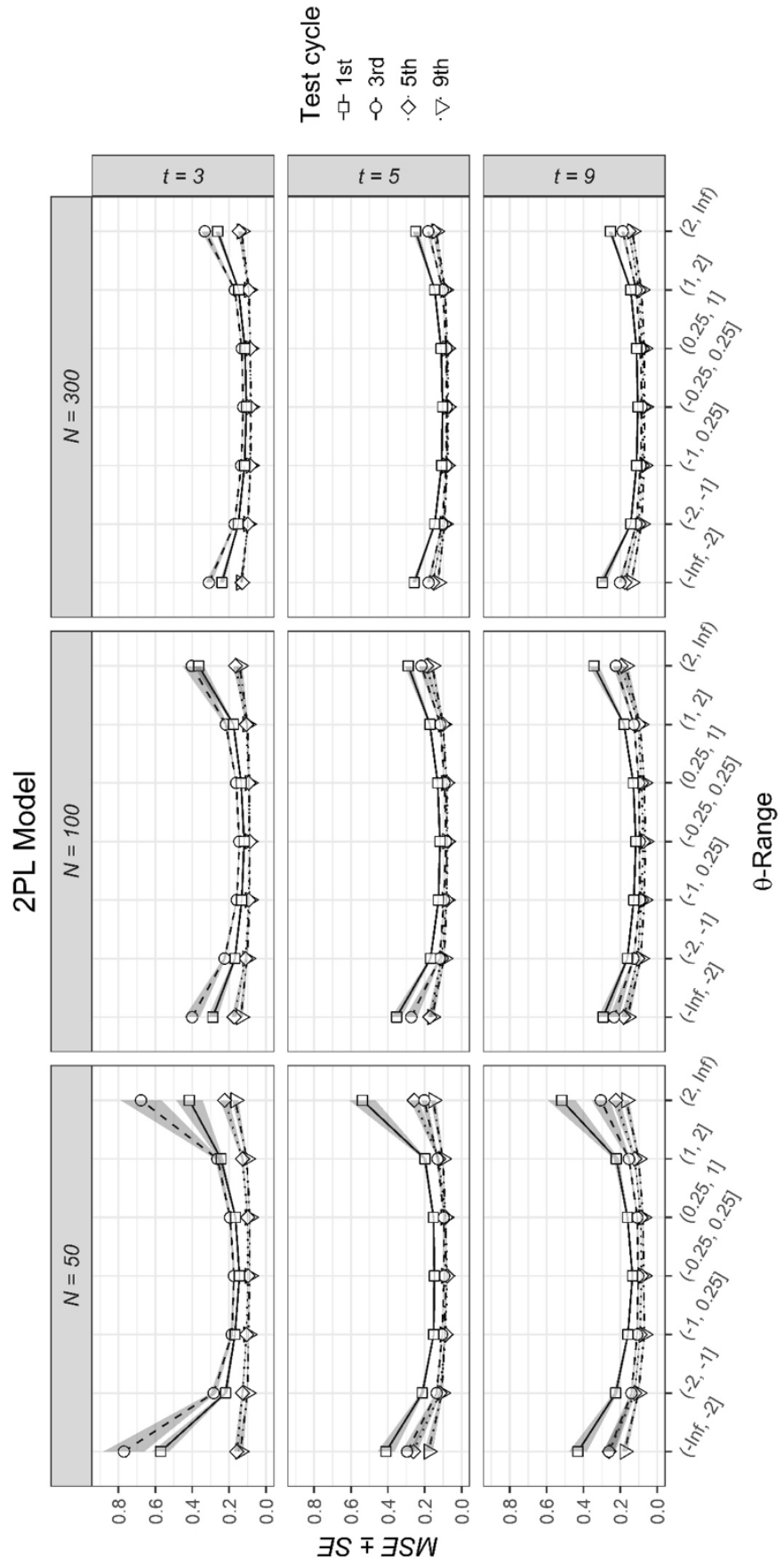


Figure 5:
Conditional mean squared error (*MSE*) at specific areas on the θ-continuum after different numbers of test cycles in the continuous calibration strategy for different sample sizes and different levels of calibration speed under the 2PL model. The area shaded gray represents the standard error around the calculated MSE obtained from the variability of all simulees in a specific θ-range.

Discussion

The general purpose of this study was to propose and examine a new continuous calibration strategy for the application of CAT in areas, where it is infeasible or difficult to conduct a calibration study before the operational phase of a computerized adaptive test and it is tolerable that measurement precision and adaptivity increase across test cycles. Such a method can be used immediately during everyday testing practice without the need for a separate calibration study. The basic ideas of the strategy are (a) constantly updating item parameters across several test cycles by using all the information from preceding test cycles, (b) adaptive item administration of calibrated items, (c) online-calibration of new items, and (d) maintaining the scale across several test cycles through linking and, therefore, allowing comparisons of the ability measures across these test cycles.

To give some practical recommendations for the configuration of the continuous calibration strategy, a simulation study was carried out, and basic tuning parameters were varied. These were (1) the sample size per test cycle, (2) the level of calibration speed, and (3) the underlying IRT model. Regarding the sample size, the results show a promising performance of the proposed strategy even for very small samples of $N = 50$. For $N = 300$ there is nearly no difference between the upper and lower baselines and the precision obtained with the continuous calibration strategy. Thus, for sample sizes such as this, it is not necessary to update the item parameters over test cycles. Nonetheless, also in this case, a test for IPD should be implemented.

With respect to the calibration speed, when using the 1PL, there is no clear recommendation. Test developers have to weigh the costs and the benefits of either a slow or a fast initial calibration. If it is tolerable that there is a substantial increase in measurement precision and if there are enough resources to construct many items for the first few test cycles, it is certainly possible to calibrate many items as fast as possible. In contrast, for the 2PL model, a slow calibration clearly outperformed a fast calibration regarding global and conditional measurement precision and therefore is recommended especially for very small samples of $N = 50$. The differences between the $t = 5$ and the $t = 9$ conditions were very small. Thus, depending on the resources available for producing new items, test developers should decide between a medium and a low level of calibration speed. A fast calibration under the 2PL model, especially in small samples, is not recommended. The selection algorithm in the 2PL model tends to select items with high discrimination parameters, which is not desired at early stages of the continuous calibration process because of larger calibration errors. For example, a good item could receive a small discrimination parameter by chance due to only a few inconsistent responses in a small sample. Therefore, this item might never get updated sufficiently, as it is simply not administered sufficiently often. As shown by van der Linden and Glas (2000), capitalization on calibration errors strongly impacts the ability estimation using the 2PL model. Thus, enlarging the item pool slowly leads to a more uniform item administration, as in the first few test cycles there are only few items the selection algorithm can choose from, which in turn leads to a higher number of item responses and more precise ability estimation.

In addition to the factors considered in this study, the proposed strategy has several interesting operational characteristics and tuning parameters that future research should address

to investigate the performance of the proposed strategy in more detail and give practical recommendations. Some of them are, for example, the test length, the method used to test for IPD, the scale transformation method, the link item selection method, and the proportion of link items. Additionally, this study was carried out under the assumption that all items follow the respective model. Future research might also investigate the impact of item misfit on the performance of the continuous calibration strategy. Therefore, methods for assessing fit in items during the calibration process should be implemented. Finally, the implemented cluster structure provides the possibility of considering item position effects during the calibration process (e.g., Frey, Bernhardt, & Born, 2017). For this purpose, the three item clusters could be divided into more fine-grained, equal-sized subclusters and presented in a balanced way across different positions.

The major conclusion that can be drawn from the results is that the continuous calibration strategy works reasonably well, even for very small samples of $N = 50$. The proposed strategy offers a practical and less resource-consuming method for test developers from application areas with the characteristics mentioned above to take advantage of the benefits of CAT.

Acknowledgements

The research reported in the article was supported by a grant from the German Federal Ministry of Education and Research (Ref: 16DHL1005).

References

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147–162. doi: 10.1111/j.1745-3984.1991.tb00350.x

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika*, *46*, 443–459. doi: 10.1007/BF02293801

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275–285. doi: 10.1111/j.1745–3984.1988.tb00308.x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48(6),* 1–29. doi:10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71(5)*, 1–39. doi:10.18637/jss.v071.i05

Chen, P. (2017). A comparative study of online calibration methods in multidimensional computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 42,* 559–590. doi: 10.3102/1076998617695098

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford.

Dolan, R. P., & Burling, K. S. (2012). Computer-based testing in higher education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (pp. 312–335). New York, NY: Routledge. doi: 10.4324/9780203142189.ch22

Frey, A. (2012). Adaptives Testen [Adaptive testing]. In H. Moosbrugger & A. Kelava (Eds.), Testtheorie und Fragebogenkonstruktion (2nd ed., pp. 275–293). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-20072-4_11

Frey, A., Bernhardt, R., & Born, S. (2017). Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests [Accounting for item position effects in the development of computerized adaptive tests]. *Diagnostica*, *63*, 167–178. doi: 10.1026/0012-1924/a000173

Frey, A., & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards [Hypothetical usage of adaptive testing for the examination of educational standards]. *Zeitschrift für Erziehungswissenschaft*, *Sonderheft 8*, 169–184. doi: 10.1007/978-3-531-90865-6_10

Goldstein, H. (1983). Measuring changes in educational attainment over time: problems and possibilities. *Journal of Educational Measurement, 20*, 369–377. doi: 10.1111/j.1745-3984.1983.tb00214.x

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144–149. doi: 10.4992/psycholres1954.22.144

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26,* 3–24. doi: 10.1177/0146621602026001001

He, W., & Reckase, M. D. (2014). Item pool design for an operational variable-length computerized adaptive test. *Educational and Psychological Measurement*, *74*, 473–494. doi: 10.1177/0013164413509629

Kingsbury, G. G. (2009). Adaptive item calibration: A simple process for estimating item parameters within a computerized adaptive test. *GMAC conference on computerized adaptive testing*. Minneapolis, MN.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer. doi: 10.1007/978-1-4939-0317-7_10

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 179–193. doi: 10.1111/j.1745-3984.1980.tb00825.x

Makransky, G., & Glas, C. A. W. (2010). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology, 11*, 1–20.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160. doi: 10.1111/j.1745-3984.1977.tb00033.x

Park, Y. S., Lee, Y.-S., & Xing, K. (2016). Investigating the impact of item parameter drift for item response theory models with mixture distributions. *Frontiers in Psychology, 255*(7), 1–17. doi: 10.3389/fpsyg.2016.00255

Patz, R. J., & Junker B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods in item response models. *Journal of educational and behavioral Statistics*, *24*, 146–178. doi: 10.2307/1165199

R Core Team (2017). *R: A language and environment for statistical computing* [Software]. R Foundation for Statistical Computing. Available from www.r-project.org

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 429–438). Boston: Elsevier Academic. doi: 10.1016/b0-12-369398-5/00444-8

Spoden, C., Frey, A. & Bernhardt, R. (in press). Running a CAT development within 18 months. *Journal of Computerized Adaptive Testing*.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29,* 6–26. doi: 10.1016/S0361-476X(02)00063-2

Stocking, M. L. (1988). Scale drift in online calibration. *ETS Research Report Series 1988*(1), 1–122. doi:10.1002/j.2330-8516.1988.tb00284.x

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210. doi:10.1177/014662168300700208

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16(1).* Retrieved from www.pareonline.net/getvn.asp?v=16&n=1.

van der Linden, W. J. (2016). *Handbook of item response theory, volume one: models*. London: Chapman and Hall.

van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 13,* 35–53. doi: 10.1207/s15324818ame1301_2

van der Linden, W. J., & Glas, C. A. W. (Eds.) (2010). *Elements of adaptive testing*. New York: Springer. doi: 10.1007/978-0-387-85461-8

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450. doi: 10.1007/bf02294627

Wells, C. S., Hambleton, R. K., Kirkpatrick, R., & Meng, Y. (2014). An examination of two procedures for identifying consequential item parameter drift. *Applied Measurement in Education, 27*, 214–231. doi: 10.1080/08957347.2014.905786

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*, 77–87. doi: 10.1177/0146621602261005

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364. doi: 10.1177/014662168400800312

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1–17. doi: 10.1207/s15326977ea1001_1

Anhang B: Beitrag 2 – Evaluating different equating setups in the continuous item pool calibration for computerized adaptive testing

Zitation: Born, S., Fink, A., Spoden, C. & Frey, A. (2019). Evaluating different equating setups in the continuous item pool calibration for computerized adaptive testing. *Frontiers in Psychology*, *10*, 1277. https://doi.org/10.3389/fpsyg.2019.01277

frontiers in Psychology

ORIGINAL RESEARCH
published: 06 June 2019
doi: 10.3389/fpsyg.2019.01277

Evaluating Different Equating Setups in the Continuous Item Pool Calibration for Computerized Adaptive Testing

Sebastian Born[1], Aron Fink[2], Christian Spoden[3] and Andreas Frey[2,4]*

[1] Department of Research Methods in Education, Institute of Educational Science, Friedrich Schiller University Jena, Jena, Germany, [2] Educational Psychology: Measurement, Evaluation and Counseling, Institute of Psychology, Goethe University Frankfurt, Frankfurt, Germany, [3] German Institute for Adult Education, Leibniz Centre for Lifelong Learning, Bonn, Germany, [4] Faculty of Educational Sciences, Centre for Educational Measurement, University of Oslo, Oslo, Norway

The increasing digitalization in the field of psychological and educational testing opens up new opportunities to innovate assessments in many respects (e.g., new item formats, flexible test assembly, efficient data handling). In particular, computerized adaptive testing provides the opportunity to make tests more individualized and more efficient. The newly developed continuous calibration strategy (CCS) from Fink et al. (2018) makes it possible to construct computerized adaptive tests in application areas where separate calibration studies are not feasible. Due to the goal of reporting on a common metric across test cycles, the equating is crucial for the CCS. The quality of the equating depends on the common items selected and the scale transformation method applied. Given the novelty of the CCS, the aim of the study was to evaluate different equating setups in the CCS and to derive practical recommendations. The impact of different equating setups on the precision of item parameter estimates and on the quality of the equating was examined in a Monte Carlo simulation, based on a fully crossed design with the factors common item difficulty distribution (bimodal, normal, uniform), scale transformation method (mean/mean, mean/sigma, Haebara, Stocking-Lord), and sample size per test cycle (50, 100, 300). The quality of the equating was operationalized by three criteria (proportion of feasible equatings, proportion of drifted items, and error of transformation constants). The precision of the item parameter estimates increased with increasing sample size per test cycle, but no substantial difference was found with respect to the common item difficulty distribution and the scale transformation method. With regard to the feasibility of the equatings, no differences were found for the different scale transformation methods. However, when using the moment methods (mean/mean, mean/sigma), quite extreme levels of error for the transformation constants *A* and *B* occurred. Among the characteristic curve method the performance of the Stocking-Lord method was slightly better than for the Haebara method. Thus, while no clear recommendation can be made with regard to the common item difficulty distribution, the characteristic curve methods turned out to be the most favorable scale transformation methods within the CCS.

Keywords: computerized adaptive test, item response theory, equating, continuous calibration, simulation

OPEN ACCESS

Edited by:
Ronny Scherer,
University of Oslo, Norway

Reviewed by:
Alvaro J. Arce-Ferrer,
Pearson, United States
Alexander Robitzsch,
Christian-Albrechts-Universität zu Kiel, Germany

****Correspondence:***
Sebastian Born
sebastian.born@uni-jena.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *23 November 2018*
Accepted: *15 May 2019*
Published: *06 June 2019*

Citation:
Born S, Fink A, Spoden C and Frey A (2019) Evaluating Different Equating Setups in the Continuous Item Pool Calibration for Computerized Adaptive Testing. Front. Psychol. 10:1277. doi: 10.3389/fpsyg.2019.01277

INTRODUCTION

The shift to using digital technology (e.g., laptops, tablets, and smartphones) for psychological and educational assessments provides the opportunity to implement computer-based state-of-the-art methods from psychometrics and educational measurement in day-to-day testing practice. In particular, computerized adaptive testing (CAT) has the potential to make tests more individualized and to enhance efficiency (e.g., Segall, 2005). CAT is a method of test assembly that uses the responses given to previously presented items for the selection of the next item (e.g., van der Linden, 2016), whereby the item that satisfies a statistical optimality criterion best is selected from a precalibrated item pool. Therefore, the calibrated item pool is an essential and important building block in CAT (e.g., Thompson and Weiss, 2011; He and Reckase, 2014). A set of items is called a calibrated item pool if the item characteristics, such as item difficulty and item discrimination, were estimated on the basis of an item response theory (IRT; e.g., van der Linden, 2016) model beforehand. However, in some contexts, such as higher education, clinical diagnosis, or personnel selection, the item pool calibration for CAT often poses a critical challenge because separate calibration studies are not feasible, and sample sizes are too low to allow for stable item parameter estimation.

To overcome this problem, Fink et al. (2018) proposed a continuous calibration strategy (CCS), which enables a step-by-step build-up of the item pool across several test cycles during the operational CAT phase. In the context of the CCS a test cycle is understood as the whole test procedure including steps like test assembly, test administration and analysis of test results. As the item parameter estimates of existing and new items are continuously updated within the CCS, equating is a critical factor to enable interchangeable score interpretation across test cycles. The equating procedure implemented in the CCS is based on a common-item non-equivalent group design (Kolen and Brennan, 2014) and is carried out in four steps: (1) common item selection, (2) scale transformation, (3) item parameter drift (IPD; e.g., Goldstein, 1983) detection, and (4) fixed common item parameter (FCIP; e.g., Hanson and Béguin, 2002) calibration.

In their study, Fink et al. (2018) evaluated the performance of the CCS for different factors (sample size per test cycle, calibration speed, and IRT model) with respect to the quality of the person parameter estimates. Although the results were promising, two issues remained open. First, the study of Fink et al. (2018) was conducted under ideal conditions (i.e., constant ability distribution of the examinees across test cycles). Second, despite the importance of the equating procedure in the CCS, its performance with respect to different setups of the procedure (i.e., selection of common items, scale transformation method, item drift detection) was not investigated in detail. For example, it became apparent that the CCS did not work as intended for very easy or very difficult items when using small sample sizes (i.e., 50 or 100 examinees) per test cycle. In these cases, item parameter estimates were biased due to a few inconsistent responses, with the consequence that these items were no longer selected by the adaptive algorithm in the following test cycles. Therefore, it was not possible to continuously update the item parameter estimates for these items.

Against this background, the aim of the present study was to investigate the performance of the equating procedure for different setups conducted under more realistic conditions (i.e., examinees' average abilities and variance differ between test cycles). The remainder of the article is organized as follows: First, we provide the theoretical background for the present study by introducing the underlying IRT model and by describing the CCS. Next, we discuss both the previously implemented equating procedure and alternative specifications. Then, we examine the performance of different setups of the different equating procedures in a simulation. Finally, we discuss the results and make recommendations for the implementation of the CCS.

THEORETICAL BACKGROUND

IRT Model

The IRT model used in this study was the two-parameter logistic (2PL) model (Birnbaum, 1968) for dichotomous items. The 2PL model defines the probability of a correct response $u_{ij} = 1$ of examinee $j = 1 \ldots N$ with a latent ability level θ_j to an item i by the following model, whereby a_i is the discrimination parameter and d_i is the easiness parameter of item i:

$$P\left(u_{ij} = 1|\theta_j, a_i, d_i\right) = \frac{\exp\left(a_i\theta_j + d_i\right)}{1 + \exp\left(a_i\theta_j + d_i\right)}, \quad (1)$$

In the traditional IRT metric where $a_i\theta_j + d_i = a_i\left(\theta_j - b_i\right)$, the a_i parameters will be the identical for these parametrizations, while the item difficulty parameter b_i is calculated as $b_i = -d_i / a_i$.

Continuous Calibration Strategy

In the following paragraphs, we briefly outline the CCS as introduced by Fink et al. (2018) and detail the equating procedure implemented. The CCS consists of two phases, a non-adaptive *initial phase* and a partly adaptive *continuous phase*. In the initial phase, which is the first test cycle of the CCS, the same items are presented to all examinees and only the item order can vary between examinees. In the continuous phase, the tests assembled consist of three types of item clusters (calibration cluster, linking cluster, adaptive cluster), whereby a cluster is comprised of several items. Each type of cluster has a specific goal. The calibration cluster offers the opportunity to include new items in the existing item pool, the linking cluster utilizes common items to allow a scale to be established across test cycles, and the adaptive cluster aims at the enhancement of measurement precision. The items in the calibration and the linking clusters are the same for all examinees and are administered sequentially, whereas the items in the adaptive cluster can differ between examinees due to the adaptive selection algorithm. Each test cycle in the continuous phase can be broken down into seven steps: (1) common item selection for the linking cluster, (2) test assembly and test administration, (3) temporary item parameter estimation, (4) scale transformation of the common items, (5)

IPD detection for the common items, (6) FCIP calibration, and (7) person parameter estimation. The equating procedure consists of four of these steps, which will be detailed in the following four paragraphs. The first three steps of the equating procedure serve as quality assurance of the common items to ensure feasible equating in the fourth step.

In the *common item selection*, items that have already been calibrated in the previous test cycles are selected as common items for the linking cluster. To ensure that the common items represent the statistical characteristics of the item pool (Kolen and Brennan, 2014), such as the range of the item difficulty, the items are assigned to five categories (very low, low, medium, high, and very high) based on their easiness parameters d_i. Fink et al. (2018) selected the items from the categories in such a way that the difficulty distribution of the common items corresponded approximately to a normal distribution. Beside the representation of the statistical item pool characteristics it is important that the common items adequately reflect the content of the item pool. This can be done by using content balancing approaches (e.g., van der Linden and Reese, 1998; Cheng and Chang, 2009; Born and Frey, 2017) within the common item selection and within the adaptive cluster.

After test assembly and test administration, the parameters for the common items are estimated based on the responses of the current test cycle. In the second step of the equating procedure, a *scale transformation* of the common items has to be conducted, because the ability distribution of the examinees usually differs between test cycles and, therefore, the item parameter estimates obtained are not directly comparable across cycles. The comparability of the parameter estimates is a necessary condition to check whether the common items are affected by IPD. For this reason, scale transformation methods (e.g., Marco, 1977; Haebara, 1980; Loyd and Hoover, 1980; Stocking and Lord, 1983) are important for the equating procedure. Fink et al. (2018) used the mean/mean method (Loyd and Hoover, 1980) for the scale transformation.

As IPD of item parameters may have a serious impact on equating results such as scaled scores and passing rates (Hu et al., 2008; Miller and Fitzpatrick, 2009), the *IPD detection* as the third step of the equating procedure is important if the method is to operate optimally. A number of tests for IPD can be used in IRT-based equating procedures, such as the Lord's χ^2-test (Lord, 1980) and the likelihood-ratio test (Thissen et al., 1988). In an iterative process of scale transformation and testing for IPD, common items that show significant IPD are excluded from the final set of common items. The iterative purification continues as long as at least one of the remaining common items shows significant IPD or less than two common items are left. The rationale behind the latter stopping rule is that at least two link items are necessary to keep the scale comparable across test cycles. Nevertheless, it should be mentioned that with a smaller number of link items, the equating procedure is more prone to sampling errors (Wingersky and Lord, 1984). Fink et al. (2018) used a one-sided *t*-test to examine whether the parameter estimates of a common item from the current test cycle differed significantly from the parameter estimates of the same item from the preceding test cycle.

The last step of the equating procedure, the *FCIP calibration*, involves the parameter estimation of all items using marginal maximum likelihood (MML; Bock and Aitkin, 1981) based on the responses from all test cycles. Because one aim of the CCS is to maintain the original scale from the initial calibration (first test cycle), the use of one step procedures (e.g., concurrent calibration; Wingersky and Lord, 1984) for estimating all item parameters of the different test cycles in one run is not suitable. If maintaining the scale from the initial calibration over the following test cycles has no priority, promising methods exist for equating multiple test forms simultaneously (Battauz, 2018). In the FCIP calibration, the parameters of the final common items are fixed at the item parameters estimated from the previous test cycle, whereas all the other items are estimated freely. If a "breakdown" occurs, which means that less than two common items remain after the IPD detection, a concurrent calibration (Wingersky and Lord, 1984) is used to establish a new scale.

Specifications of the Common Item Selection

The common item selection and the scale transformation of the common items are crucial parts of the CCS because they ensure that the procedure functions well. In terms of the common item selection, different distributional assumptions such as an approximated normal distribution, as used in Fink et al. (2018), or a uniform distribution may underlie the item selection. Up to now, only Vale et al. (1981) examined the impact of different common item distributions on the accuracy of the item parameter estimates using the mean/sigma method (Marco, 1977). The authors selected the common items in such a way that the test information curves of the common items were peaked (with the most information at theta equals zero) or had an approximately normal or uniform shape. In terms of the bias of the item parameter estimates, the peaked test information curve performed worst. There were only slight differences in the performance, depending on whether normally or uniformly shaped test information curves were used for the common items. As an alternative, items with extreme difficulties (bimodal distribution) might be selected as common items for the linking cluster and, therefore, might be administered to all examinees. As a consequence, the number of responses for these items increases and the impact of the few inconsistent responses that might cause bias in the estimates and prevent later administration and parameter updating in the following test cycles would be reduced. Because the quality of the equating highly depends on the common items selected, it may be argued that especially a bimodal distribution of the common items threatens the goal of maintaining the scale across test cycles. However, the item drift test implemented in the CCS ensures that significant changes in the parameter estimates of the common items between test cycles do not affect the later FCIP calibration that is used to maintain the scale.

Scale Transformation

When item parameters are estimated using different groups of examinees, the obtained parameters are often not comparable

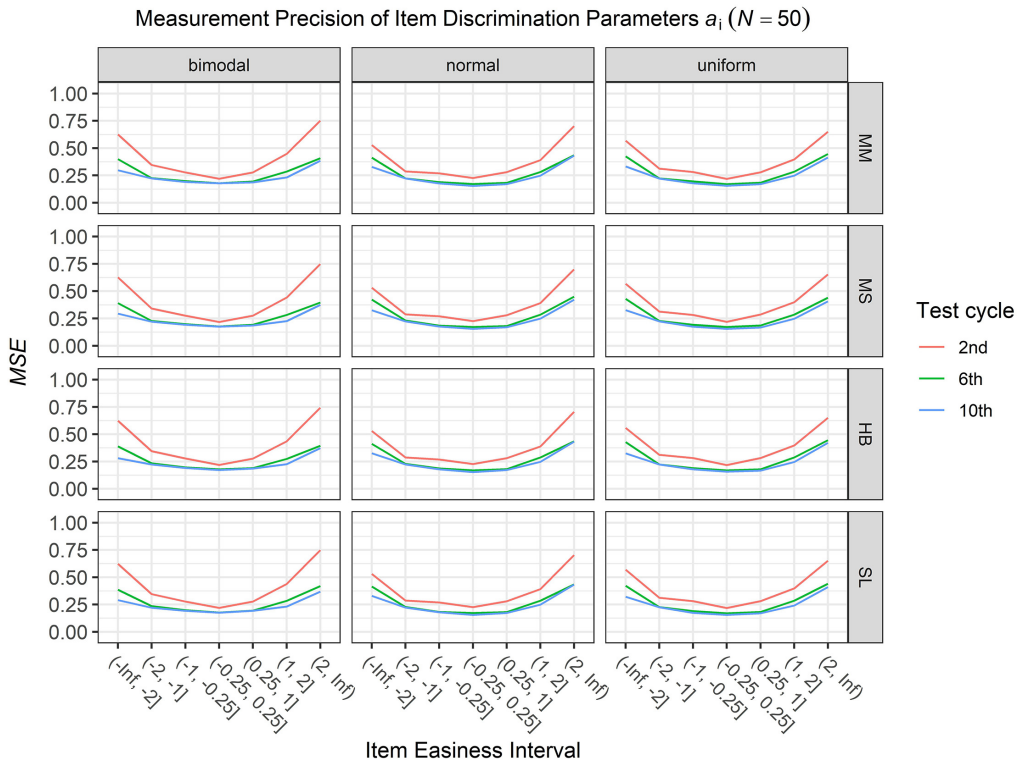


FIGURE 1 | Conditional mean squared error (*MSE*) of the item discrimination a_i for specific item easiness intervals after the 2nd, 6th, and 10th test cycles in the continuous calibration strategy with a sample size per test cycle of $N = 50$ for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

due to arbitrary decisions that have been made to fix the scale of the item and person parameter space (Yousfi and Böhme, 2012). In that case, the comparability of the item parameters can be attained by an IRT scale transformation. If the underlying IRT model holds for two groups of examinees, K and L, then the logistic IRT scales differ by a linear transformation for both the item parameters and the person parameters (Kolen and Brennan, 2014). The linear equation for the θ-values can be formulated as follows:

$$\theta_{Lj} = A\theta_{Kj} + B, \tag{2}$$

where A and B represent the transformation constants (also referred to as slope and shift) and θ_{Kj} and θ_{Lj} the person parameter values for an examinee j on scale K and scale L. The item parameters for the 2PL model on the two scales are defined in Eqs 3 and 4, where a_{Ki}, b_{Ki}, and a_{Li}, b_{Li} represent the item parameters on scale K and on scale L, respectively.

$$a_{Li} = \frac{a_{Ki}}{A} \tag{3}$$

$$b_{Li} = Ab_{Ki} + B \tag{4}$$

To obtain the transformation constants A and B, several scale transformation methods can be used. The *moment methods* such as the mean/mean and the mean/sigma express the relationship of scales by using the means and standard deviations of item or person parameters, whereas the *characteristic curve methods* minimize a discrepancy function with respect to the item characteristic curves (Haebara, 1980) or the test characteristic curve (Stocking and Lord, 1983). Research comparing these methods has found that characteristic curve methods produced more stable results compared to the moment methods (e.g., Baker and Al-Karni, 1991; Kim and Cohen, 1992; Hanson and Béguin, 2002). Within the moment methods, the mean/mean method turned out to be more stable (Ogasawara, 2000). Furthermore, Kaskowitz and de Ayala (2001) found that characteristic curve methods were robust against moderate estimation errors and were more accurate with a larger number of common items (15 or 25 compared to only five common items). In sum, the moment methods are easily implementable, but the characteristic curve methods seem to be more robust against estimation errors.

RESEARCH QUESTIONS

As the purpose of equating procedures in the CCS is to enable an interchangeable score interpretation across test cycles, the selection of the common items is a crucial factor for feasible equating. Up to now, only recommendations for the number of common items that should be used when conducting IRT equating have been made (Kolen and Brennan, 2014). Furthermore, it is suggested that the common items should

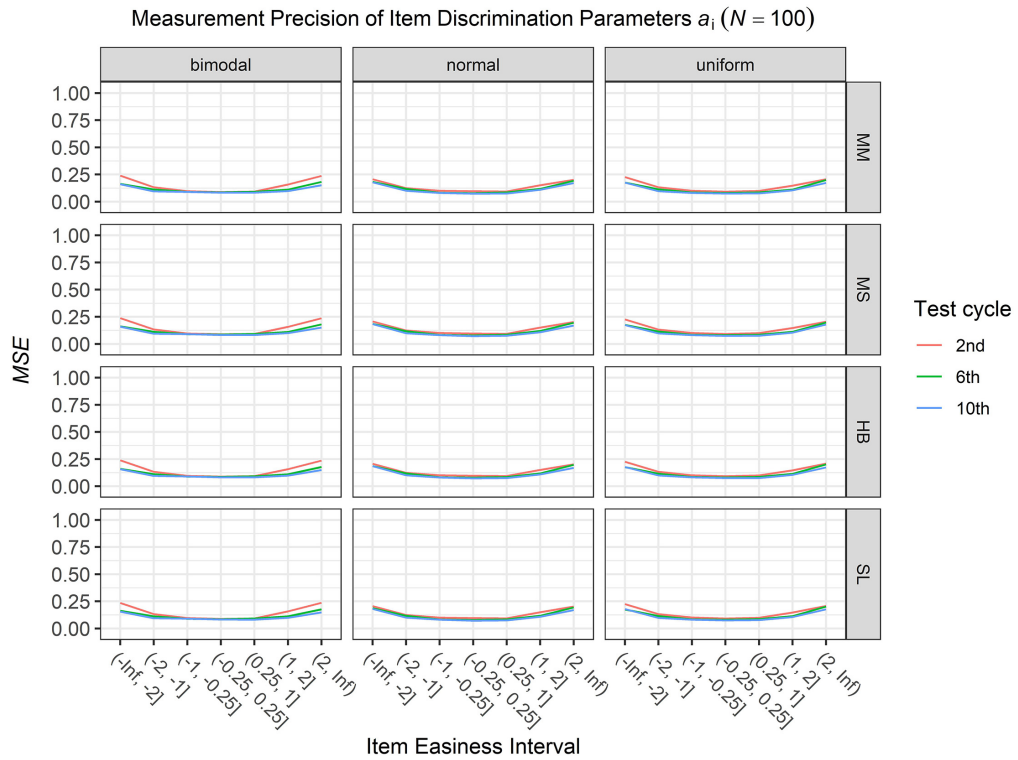


FIGURE 2 | Conditional mean squared error (*MSE*) of the item discrimination a_i for specific item easiness intervals after the 2nd, 6th, and 10th test cycle in the continuous calibration strategy with a sample size per test cycle of N = 100 for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

represent the content and statistical characteristics of the test or rather the complete item pool. For example, modifying the common item selection in such a way that more items with extreme item difficulty levels are included may enhance the precision of these items, but it could threaten the quality of the equating. Therefore, our first two research questions can be formulated as follows:

1. What effect does the difficulty distribution of the common items in the CCS have on the precision of the item parameter estimates?
2. What effect does the difficulty distribution of the common items in the CCS have on the quality of the equating?

Fink et al. (2018) used the mean/mean method for scale transformation because of its simple and user-friendly implementation. Given prior research on scale transformation methods, this might not be the best choice when the sample size per test cycle is low. Furthermore, there are several packages for the open-source software R (R Core Team, 2018) available to implement the characteristic curve methods (e.g., Weeks, 2010; Battauz, 2015). As already mentioned above, the scale transformation method used and the IPD detection implemented in the CCS could serve as quality assurance to ensure that significant changes in the parameter estimates of the common items between test cycles do not affect the later FCIP calibration. For this reason, our third research question is:

3. What effect does the scale transformation method used in the CCS have on the quality of the equating?

As the CCS was developed for a context in which separate calibration studies are often not feasible and sample sizes are too low to allow for stable item parameter estimation, it is important to evaluate whether the results for these three research questions were affected by the sample size. Consequently, each of the three research questions was investigated with a special focus on additional variations of the sample size.

MATERIALS AND METHODS

Study Design

Many factors can affect the quality of the equating within the CCS. These include, among others, the number of common items, the test length, the characteristics of the common items, the scale transformation method applied, the number of examinees per test cycle, the presence of IPD and the test applied for IPD. In the present study, some of these factors were kept constant (e.g., number of common items, test length, the presence of IPD, test applied for IPD) to ensure the comprehensibility of the study results.

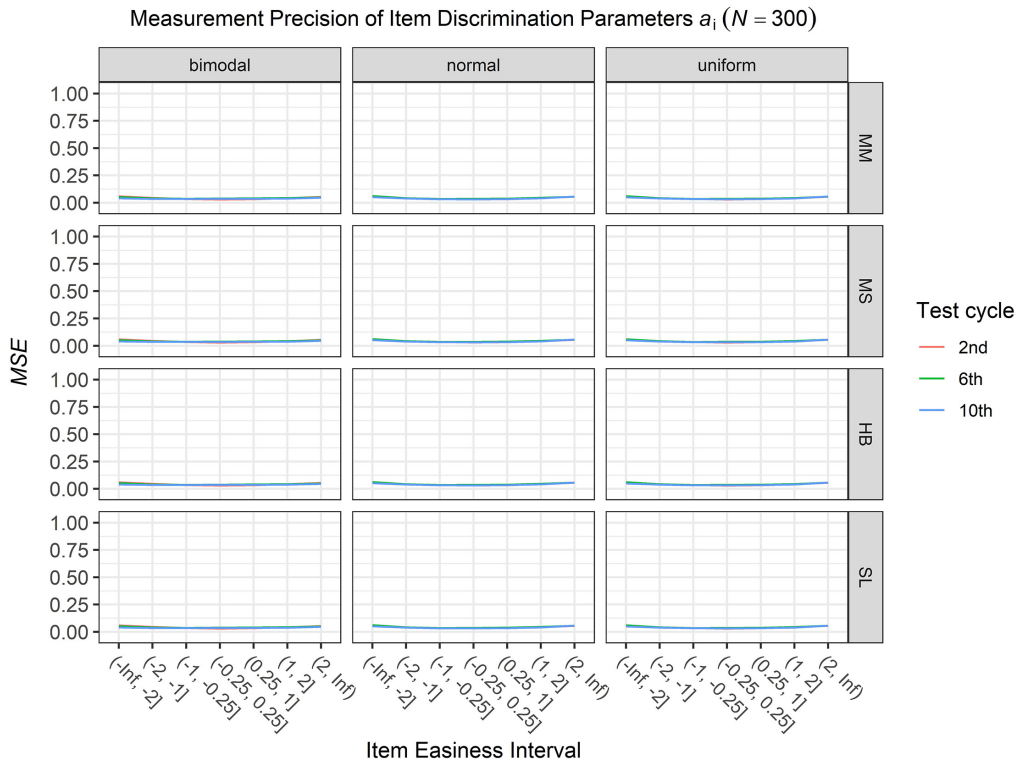


FIGURE 3 | Conditional mean squared error (*MSE*) of the item discrimination a_i for specific item easiness intervals after the 2nd, 6th, and 10th test cycle in the continuous calibration strategy with a sample size per test cycle of N = 300 for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

To answer the research questions stated above, a Monte Carlo simulation based on a full factorial design with three independent variables (IVs) was conducted. With the first IV, *difficulty distribution*, the distribution of easiness parameters d_i of the common items (normal, uniform, and bimodal with very low and very high difficulties only) was varied. The second IV, *transformation method*, compared the most common scale transformation methods (mean/mean, mean/sigma, Haebara, and Stocking-Lord) used for computing the transformation constants to conduct the scale transformation. The third IV, *sample size,* reflected the number of test takers per test cycle (N = 50; N = 100; N = 300). Because the CCS uses the responses from multiple test cycles, the number of test takers per test cycle chosen for the study is small compared to the recommendations (e.g., a minimum of 500 responses per item for the 2PL model; de Ayala, 2009). The fully crossed design comprised $3 \times 4 \times 3 = 36$ conditions. For each of the conditions, 200 replications were conducted and analyzed with regard to various evaluation criteria (see below).

The simulations were carried out in R (R Core Team, 2018) using the "mirtCAT" package (Chalmers, 2016) for simulating adaptive tests and the "mirt" package (Chalmers, 2012) for item and person parameter estimation. Transformation constants were calculated based on the common items of consecutive test cycles using the "equateIRT" package (Battauz, 2015). The test for IPD was also conducted with the "equateIRT" package. We decided to use the "equateIRT" package in the simulations because it enables a direct import of results from the "mirt" package and offers an implemented test for IPD. The corresponding functions were called in a R script, which was written to carry out the CCS.

Simulation Procedure

Data Generation

In each replication, the discrimination parameters a_i were drawn from a lognormal distribution, $a_i \sim \log N\,(0,\ 0.25)$, and the easiness parameters d_i were drawn from a truncated normal distribution, $d_i \sim N\,(0,\ 1.5)$, $d_i \in (-2.5,\ 2.5)$. Since this study was not designed to investigate IPD detection rates (e.g., Battauz, 2019), no IPD was simulated in the data. Therefore the true item parameters a_i and d_i remained unchanged over the test cycles.

The ability parameters of the examinees in the first test cycle in each replication were randomly drawn from a standard normal distribution, $\theta \sim N\,(0,\ 1)$. For the subsequent test cycles t within a replication, the ability parameters followed a normal distribution, $\theta \sim N\,(\mu_t,\ \sigma_t)$, whereby the mean $\mu_t \in (-0.5,\ 0.0,\ 0.5)$ and the standard deviation $\sigma_t \in (0.7,\ 1.0,\ 1.3)$ were randomly drawn. This was done to mimic the fact that examinees of different test cycles usually differ with respect to the mean and variance of their ability distribution. The examinees' responses to the items were generated in line with the 2PL model.

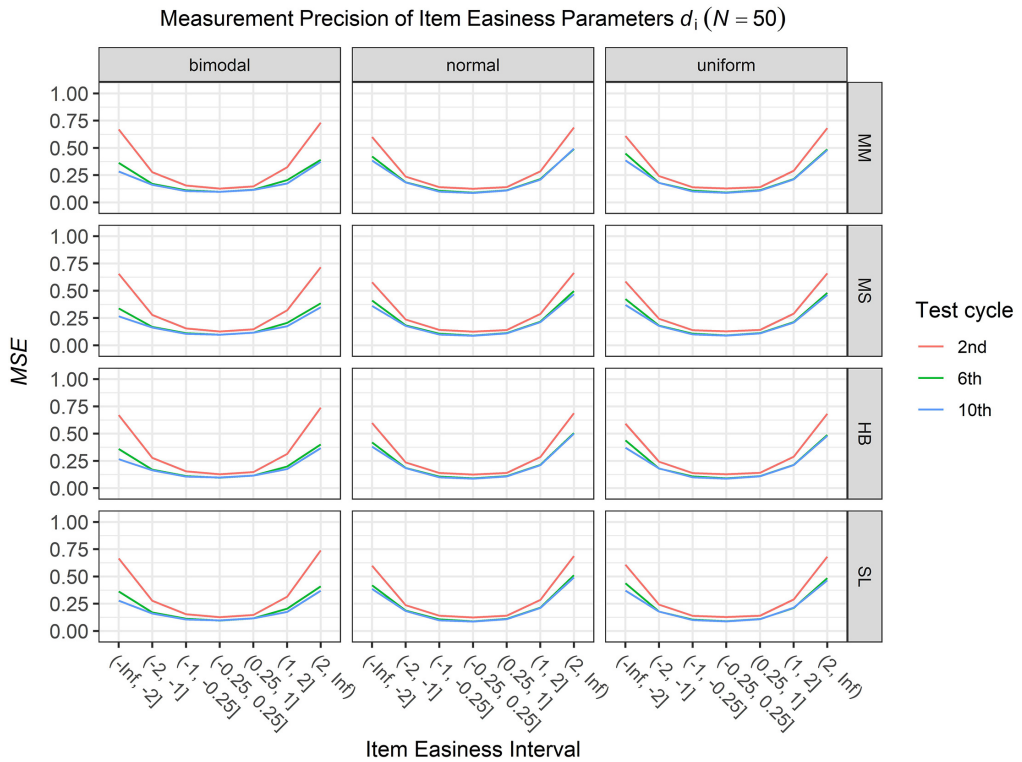


FIGURE 4 | Conditional mean squared error (*MSE*) of the item easiness d_i for specific item easiness intervals after the 2nd, 6th, and 10th test cycle in the continuous calibration strategy with a sample size per test cycle of $N = 50$ for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

Specification of the CCS

The CCS in the current study was applied with all seven steps proposed by Fink et al. (2018) including the IPD detection of the common items. Although no IPD was simulated in the data, in realistic settings the untested assumption of item parameter invariance is questionable. Even in the absence of IPD item parameters can significantly differ between test cycles because of sampling error. The number of test cycles within the CCS was set to 10 test cycles, whereby the first test cycle represented the initial phase and the subsequent test cycles the continuous phase. The test length was kept constant with 60 items. The calibration cluster in the continuous phase consisted of 20 items, resulting in an item pool size of $I_t = 60 + (t - 1) \cdot 20$ after the test cycle t, and a total item pool size of 240 items after the 10th test cycle. Following the recommendation of Kolen and Brennan (2014) that the number of common items should be at least 20% of the test length, the number of common items in the linking cluster was set to 15 items. Consequently, the adaptive cluster in each test cycle of the continuous phase contained 25 items. Within the adaptive cluster, the *maximum a posteriori* (MAP; Bock and Aitkin, 1981) was used as the ability estimator and the maximum information criterion (Lord, 1980) was applied for the adaptive item selection.

For the common item selection within the equating procedure, only items that had already been calibrated in the previous test cycles and that did not serve as common items in the preceding test cycle were eligible. The selection procedure for the common items differed depending on the intended distribution. For the normal distribution, the procedure of Fink et al. (2018) was applied. The eligible items were first assigned to five categories (very low, low, medium, high, and very high) based on their easiness parameters d_i. Then, five items from the "medium" category, three items each from the "low" and "high" categories, and two items from each of the extreme categories were chosen to mimic a normal distribution. For the uniform distribution, the eligible items were assigned to 15 categories based on their easiness parameters d_i and one item from each category was drawn. The interval limits of the categories were determined as quantiles of the item difficulty distribution. For the bimodal distribution, the eligible items were ordered according to their easiness parameters d_i and two subsamples were formed containing the 11 easiest and the 11 hardest items, respectively. Then, 15 items in total were randomly drawn from the two subsamples (seven easy and eight difficult items, or vice versa). As already mentioned, the selected common items in periodical assessments should be comparable also with regard to content characteristics. Content balancing approaches like the maximum priority index (Cheng and Chang, 2009) and the shadow testing approach (van der Linden and Reese, 1998) may be used for this purpose. Because no substantial impact was expected on the

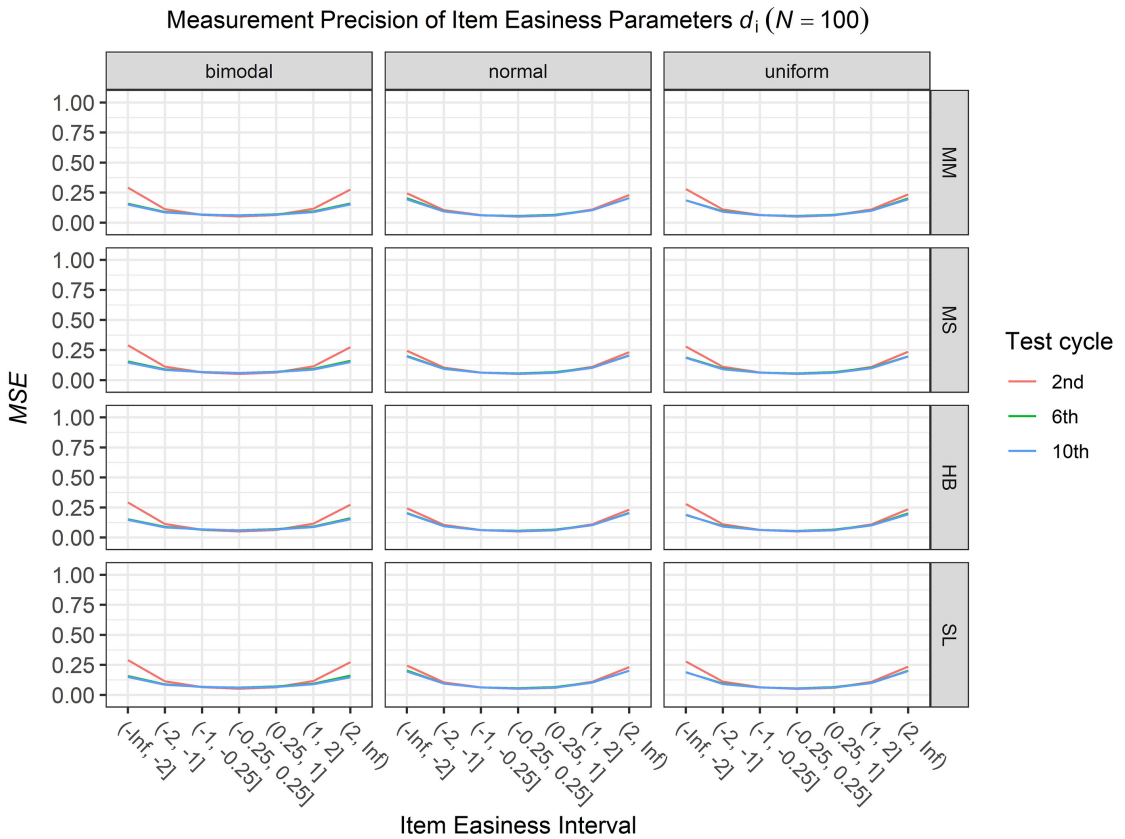


FIGURE 5 | Conditional mean squared error (*MSE*) of the item easiness d_i for specific item easiness intervals after 2nd, 6th, and 10th test cycle in the continuous calibration strategy with a sample size per test cycle of N = 100 for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

measurement precision of the item parameters or on the quality of the equating, content balancing was not considered as a factor in the study.

For the scale transformation, one of the four transformation methods (Mean/Mean, Mean/Sigma, Haebara, and Stocking-Lord) was applied. A modified version of Lord's chi-squared method (Lord, 1980) that is implemented in the "equateIRT" package (Battauz, 2015) was used as the test for IPD with a type I error level of 0.05. In an iterative purification process (Candell and Drasgow, 1988) of scale transformation and testing for IPD, items that showed significant IPD were removed from the set of common items. In each test cycle, MML estimation was used to obtain the item parameters for both the temporary item parameter estimation and the FCIP calibration. The lower and the upper bound for the item discrimination a_i was set to –1 and 5, respectively. For the item easiness parameters d_i, the bounds were set to –5 and 5.

Evaluation Criteria

The mean squared error (*MSE*) of the item parameters a_i and d_i, respectively, was calculated after each test cycle t as the averaged squared difference between the item parameter estimates and the true item parameters for all items I_t across all replications R = 200. Thus, a high degree of precision is denoted by low values for the *MSE*.

$$MSE_t(a_i) = \frac{1}{R^* I_t} \sum_{r=1}^{R} \sum_{i=1}^{I_t} \left(\hat{a}_{ir} - a_{ir}\right)^2 \quad (5)$$

$$MSE_t(d_i) = \frac{1}{R^* I_t} \sum_{r=1}^{R} \sum_{i=1}^{I_t} \left(\hat{d}_{ir} - d_{ir}\right)^2 \quad (6)$$

Because our aim was to evaluate whether the modified common item selection could prevent a dysfunction of the CCS in terms of more precise item parameter estimates for items with very low and very high values for d_i, the conditional *MSE* was used as a criterion. Therefore, the *MSE* was calculated for seven easiness intervals: $d_i \in (-Inf, -2]$, $d_i \in (-2, -1]$, $d_i \in (-1, -0.25]$, $d_i \in (-0.25, 0.25]$, $d_i \in (0.25, 1]$, $d_i \in (1, 2]$, and $d_i \in (2, Inf)$.

Three criteria were used to evaluate the equating quality. As a first criterion, we used the proportion of test cycles in which no breakdown of the common items occurred. Second, we calculated the proportion of drifted items for each of the 36 conditions. And third, we computed the accuracy (*Error*) of the scale transformation constants A and B for each replication r

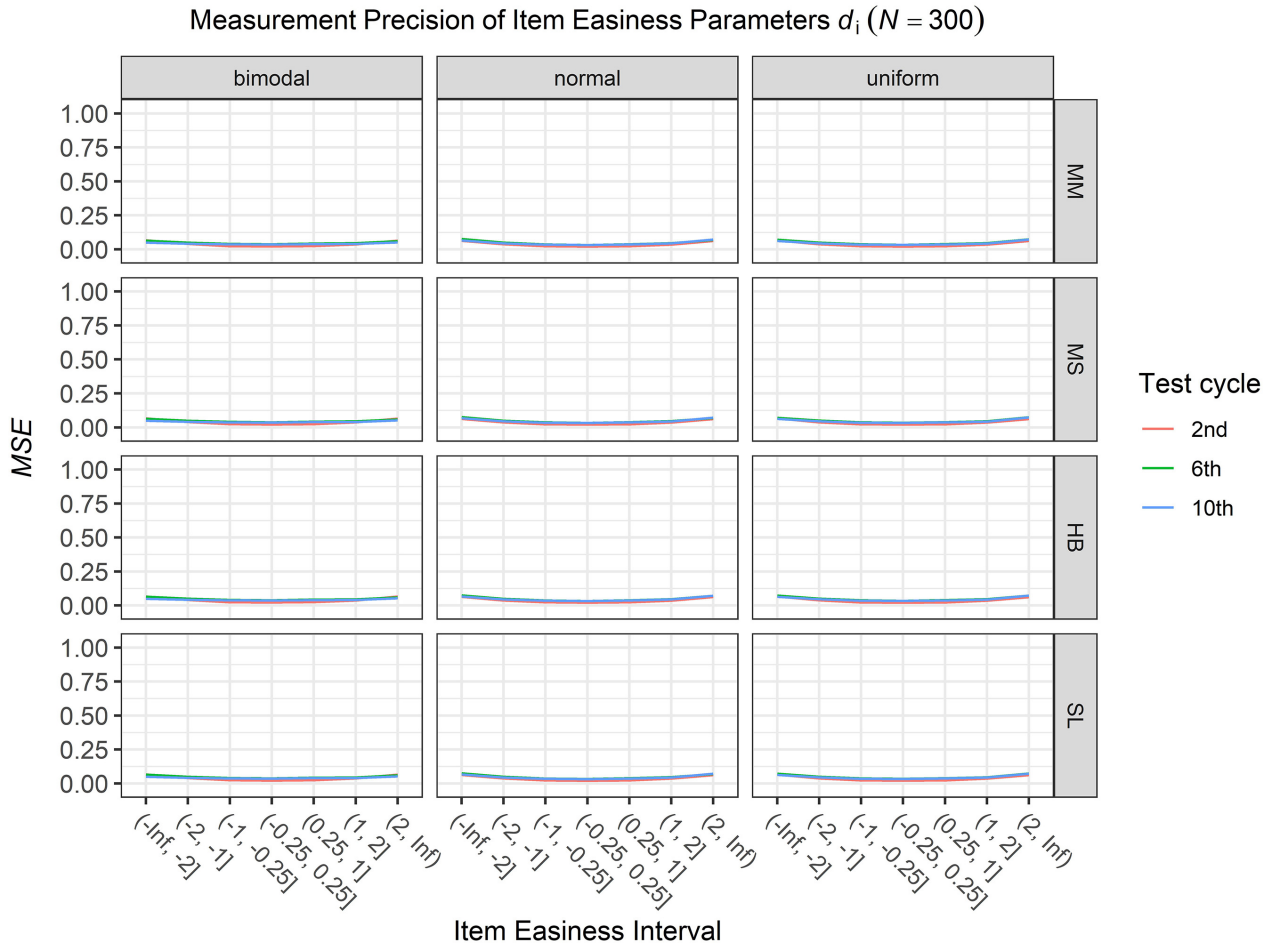


FIGURE 6 | Conditional mean squared error (*MSE*) of the item easiness d_i for specific item easiness intervals after the 2nd, 6th, and 10th test cycle in the continuous calibration strategy with a sample size per test cycle of $N = 300$ for different common item difficulty distributions and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord).

when no breakdown occurred as the difference between the true and the estimated transformation constants for every test cycle in the continuous phase. The average of the *Error* corresponds to the Bias of the transformations constants.

$$Error\left(A_{tr}\right) = \left(\hat{A}_{tr} - A_{tr}\right) \tag{7}$$

$$Error\left(B_{tr}\right) = \left(\hat{B}_{tr} - B_{tr}\right) \tag{8}$$

The true transformation constants A and B were calculated based on the true examinees' abilities from/in all previous test cycles p and from/in the current test cycle t (Kolen and Brennan, 2014).

$$A_t = \frac{\sigma\left(\theta_t\right)}{\sigma\left(\theta_p\right)} \tag{9}$$

$$B_t = \mu\left(\theta_t\right) - A\mu\left(\theta_p\right) \tag{10}$$

The estimated transformation constants $\hat{A}_t$ and $\hat{B}_t$ were obtained based on the parameter estimates of the final set of common items from the previous and the current test cycles using one of the four scale transformation methods implemented in the "equateIRT" package (Battauz, 2015). The third criterion was calculated only for the cases where at least two common items remained after the IPD detection.

RESULTS

Note that the conditions with the mean/mean method as scale transformation method and normal distributed common items mimic the setup of the equating procedure from Fink et al. (2018).

Conditional Precision of Item Parameters

To answer the first research question regarding the precision of the item parameter estimates, we analyzed the conditional *MSE* of the item discrimination parameters a_i and the item easiness parameters d_i depending on the scale transformation method, the common item difficulty distribution, and the sample sizes per test cycle. For the sake of clarity, the results are only

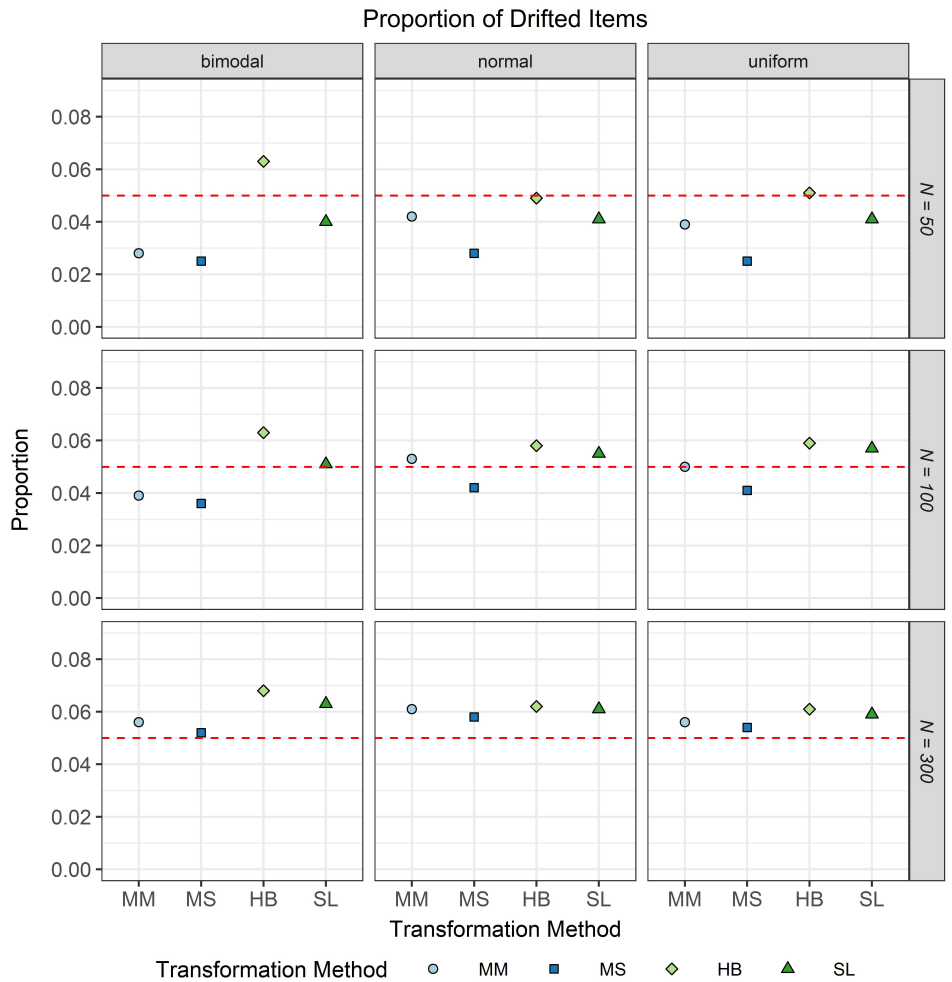


FIGURE 7 | Proportion of drifted items in the continuous calibration strategy for different sample sizes per test cycle, different common item difficulty distributions, and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord). The dashed line represents the type I error level of 0.05.

presented for the second, the sixth, and the 10th test cycles of the CCS. **Figures 1–3** illustrate the conditional *MSE* of the item discrimination parameter estimates a_i, and **Figures 4–6** illustrate the conditional *MSE* of the item easiness parameter d_i. As can be expected based on the findings from Fink et al. (2018), the *MSE* for the item discrimination parameter estimates and the item easiness parameter estimates decreased as the number of test cycles in the CCS increased and as the sample size per test cycle increased. With regard to the precision of the item parameter estimates, no substantial differences were found between the different scale transformation methods, independent of the common item difficulty distribution and the sample size per test cycle. When a bimodal difficulty distribution of common items was chosen, the precision of the item parameter estimates for the very easy and very difficult items was higher compared to a normal or uniform difficulty distribution of common items (**Figures 1**, **4**). However, this minimal gain came at the expense of a lower precision of the item parameter estimates for items with medium difficulty. This effect was found for very small sample sizes per test cycle ($N = 50$), and diminished for larger sample sizes ($N = 100$, $N = 300$).

Quality of Equating

The second and third research questions focused on the equating procedure. The first evaluation criterion was the proportion of feasible equatings (at least two items remained after the IPD detection). Most striking was that over all replications for none of the test cycles a breakdown of the common items occurred. Furthermore, for all 36 conditions the median number of eligible common items over all test cycles and replications ranged from 14 to 15.

The second evaluation criterion was the proportion of drifted items. As IPD was not simulated in the study and because the type I error level of the test for IPD was set to 0.05, it was expected that approximately five percent of the common items would show significant IPD. **Figure 7** shows the proportion of drifted common items depending on the common item difficulty distribution, the scale transformation method, and the sample

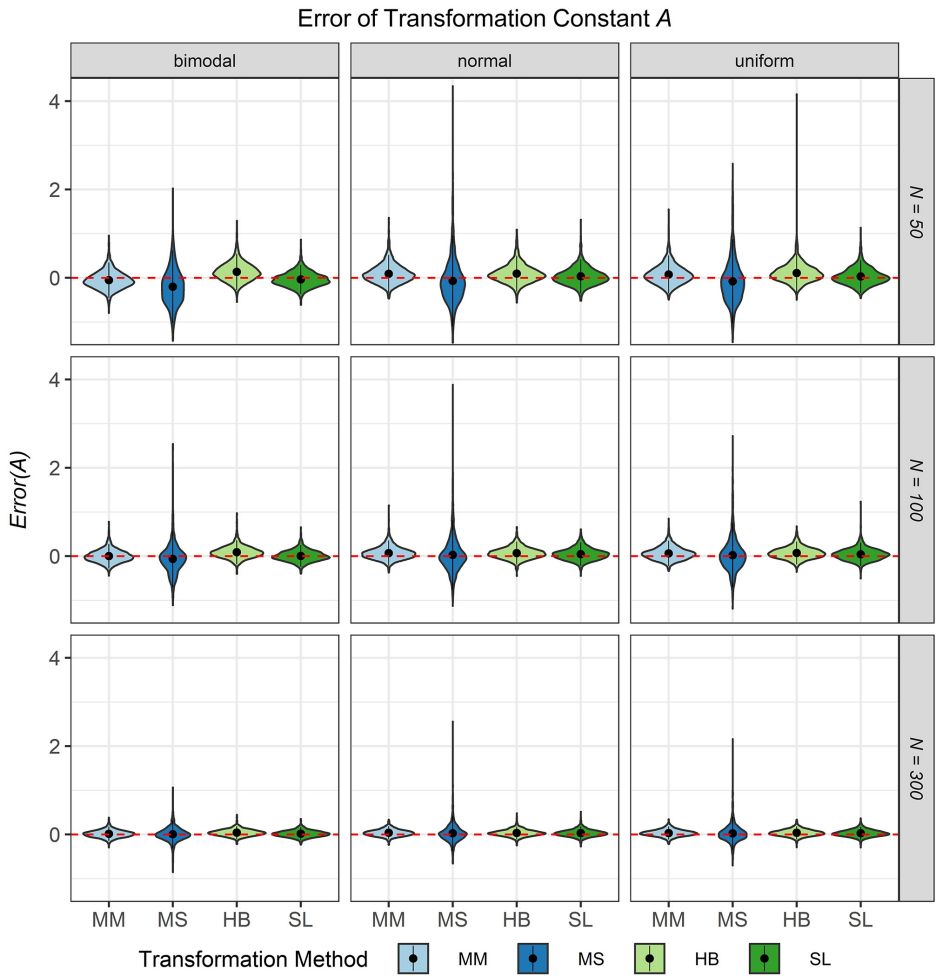


FIGURE 8 | Error of the transformation constant *A* in the continuous calibration strategy for different sample sizes per test cycle, different common item difficulty distributions, and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord). The dot in the middle of each violin represents the bias and the width of the violin expresses the frequency of the corresponding value.

size per test cycle. It is obvious from this figure that independent of the scale transformation method and the common item difficulty distribution, the type I error rates increased with increasing sample size per test cycle. This effect was stronger for the moment/methods. Furthermore, it became apparent that if the difficulty distribution of the common items was uniform or normal, all scale transformation methods did not considerably differ from the type I error level of 0.05. The only exception to this result was the mean/sigma method which generally led to considerably smaller type I error rates when the sample size was small (N = 50). All in all, using the Stocking-Lord method resulted for all conditions in type I error rates that did not considerably differ from the type I error level of 0.05.

The third evaluation criterion was the accuracy of the transformation constants A and B when no breakdown occurred. **Figures 8**, **9** show violin plots for the *Error* of the transformation constants A and B depending on the common item difficulty distribution, the scale transformation method, and the sample size per test cycle. In violin plots, the frequency distribution of a numeric variable (e.g., bias) is expressed. Note that the average error (= *Bias*; represented by the dot in the violin) for both transformation constants A and B did not differ substantially from zero for all scale transformation methods, independent of the common item difficulty distribution and the sample size per test cycle. However, the variation of the error (represented by the height of the violin) differed between the scale transformation methods and, especially for the moment methods rather high levels of error occurred. The characteristic curve methods showed the lowest variation in error. With increasing sample size per test cycle, the variation of the error decreased, but there were still extreme levels of error for the mean/mean and the mean/sigma method.

In summary and in terms of the three research questions, the study provided the following results:

1. The difficulty distribution of the common items in the CCS did not have a substantial impact on the precision of the item parameter estimates

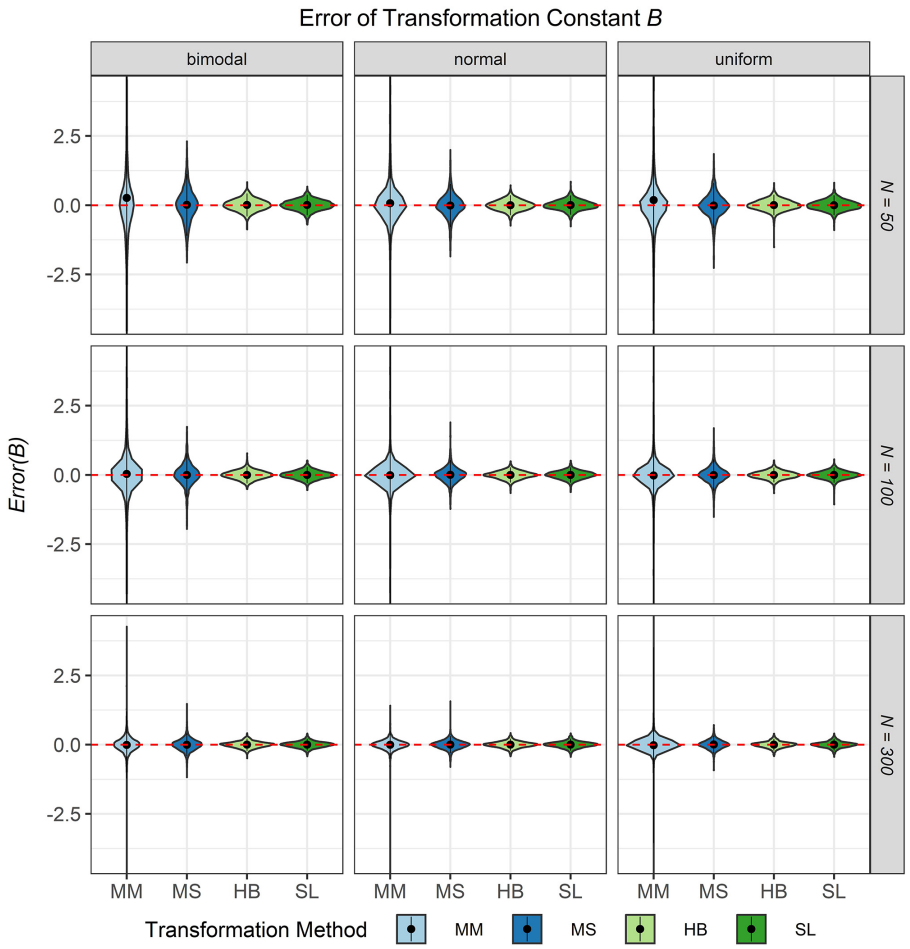


FIGURE 9 | Error of the transformation constant *B* in the continuous calibration strategy for different sample sizes per test cycle, different common item difficulty distributions, and different scale transformation methods (MM = Mean/Mean, MS = Mean/Sigma, HB = Haebara, SL = Stocking-Lord). The dot in the middle of each violin represents the bias and the width of the violin expresses the frequency of the corresponding value.

although small differences existed between the common item distributions; these differences were in opposite/varying directions for extreme and medium-ranged item easiness parameters d_i when the sample size was very small.

2. With regard to the proportion of feasible equatings (at least two common items remained after the test for IPD) no differences were found independent of the common item difficulty distributions, the scale transformation method and the sample size.
3. The characteristic curve methods outperformed the moment methods in terms of error of the transformation constant. Especially for small sample size the mean/sigma method cannot recommended.

DISCUSSION

The objective of the present study was to evaluate different setups of the equating procedure implemented in the CCS and to make/provide recommendations on how to apply these setups. For this purpose, the quality of the item parameter estimates and of the equating was examined in a Monte Carlo simulation for different common item difficulty distributions, different scale transformation methods, and different sample sizes per test cycle.

The following recommendations can be made based on the results obtained: First, no clear advantage of using any of the three common item difficulty distributions was identified. Regarding the precision of the item parameter estimates, the results show a slight increase in the precision of the item parameter estimates for items with extreme difficulties when using a bimodal common item difficulty distribution compared to a normal or uniform distribution. However, the precision of the item parameter estimates for items with medium difficulty decreased. These effects were only found for very small sample sizes per test cycle (N = 50) and no differences were found for larger sample sizes (N = 100, N = 300). Furthermore, the use of different scale transformation methods did not have a substantial effect on the precision of the item parameter estimates.

Note that exposure control methods (e.g., Sympson and Hetter, 1985; Revuelta and Ponsoda, 1998; Stocking and Lewis, 1998) might be an alternative to increase the number of responses to items with extreme difficulty levels and, in consequence, the precision of the item parameter estimates for these items. However, using these methods would sacrifice adaptivity to a certain degree and, thus, the efficiency of the computerized adaptive test (e.g., Revuelta and Ponsoda, 1998). This is even more relevant to tests assembled within the partly adaptive CCS, because only one of the three cluster types used is based on an adaptive item selection. Furthermore, in the early stages of the CCS, the item pool is rather small, which also limits the adaptivity of the tests. For these reasons, it can be expected that exposure control methods do not offer an ideal option for the CCS to increase the precision of item parameter estimates for items with extreme difficulties. This point might be examined by future research.

Second, with respect to the quality of the equating, no difference was found for the scale transformation methods with regard to the proportion of feasible equatings independent of the common item difficulty distribution used and the sample size available per test cycle. The rule for evaluating an equating as feasible (at least two common items remained after the test for IPD) is worthy of discussion because of two reasons: first, with a small number of remaining common items, the equating procedure is more prone to sampling error (Wingersky and Lord, 1984) and second, it is rather unlikely that the content of the item pool is adequately reflected by the remaining common items. However, even if the criterion for evaluating an equating as feasible had been set to ten remaining common items, the proportion of feasible equatings would be at least 99% in all conditions. With regard to the type I error rate and the error of the transformation constant the characteristic curve methods outperformed the moment methods especially for small sample sizes. This is in line with the result of Ogasawara (2002) who found that the characteristic curve methods are less affected by imprecise item parameter estimates and lead to more accurate transformation than moment methods. Among the characteristic curve methods the Stocking-Lord method was slightly better than the Haebara method in almost all conditions. Thus, although our results do not facilitate a clear recommendation regarding the most favorable common item difficulty distribution, they do enable a clear recommendation in terms of the preferred scale transformation method: The Stocking-Lord method should be used as the scale transformation method within the CCS.

AUTHOR CONTRIBUTIONS

SB conceived the study, conducted the statistical analyses, drafted the manuscript, and approved the submitted version. AFi performed substantial contribution to the conception of the study, contributed to the programming needed for the simulation study (R), reviewed the manuscript critically for important intellectual content, and approved the submitted version. CS performed substantial contributions to the interpretation of the study results, reviewed the manuscript critically for important intellectual content, and approved the submitted version. AFr provided advise in the planning phase of the study, reviewed the manuscript critically for important intellectual content, and approved the submitted version.

FUNDING

The research reported in the article was supported by a grant from the German Federal Ministry of Education and Research (Ref: 16DHL1005).

REFERENCES

Baker, F. B., and Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *J. Educ. Meas.* 28, 147–162. doi: 10.1111/j.1745-3984.1991.tb00350.x

Battauz, M. (2015). equateIRT: an R package for IRT test equating. *J. Stat. Softw.* 68, 1–22. doi: 10.18637/jss.v068.i07

Battauz, M. (2018). "Simultaneous equating of multiple forms," in *Quantitative Psychology*, eds M. Wiberg, S. Culpepper, R. Janssen, J. González, and D. Molenaar (Cham: Springer), 121–130.

Battauz, M. (2019). On wald tests for differential item functioning detection. *Stat. Methods Appl.* 28, 121–130. doi: 10.1007/s10260-018-00442-w

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, eds F. M. Lord and M. R. Novick (Reading, MA: Addison-Wesley), 395–479.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika* 46, 443–459. doi: 10.1007/BF02293801

Born, S., and Frey, A. (2017). Heuristic constraint management methods in multidimensional adaptive testing. *Educ. Psychol. Meas.* 77, 241–262. doi: 10.1177/0013164416643744

Candell, G. L., and Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Appl. Psychol. Meas.* 12, 253–260. doi: 10.1177/014662168801200304

Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *J. Stat. Softw.* 48, 1–29. doi: 10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *J. Stat. Softw.* 71, 1–39. doi: 10.18637/jss.v071.i05

Cheng, Y., and Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *Br. J. Math. Stat. Psychol.* 62, 369–383. doi: 10.1348/000711008X304376

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NL: Guilford.

Fink, A., Born, S., Spoden, C., and Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychol. Test Assess. Model.* 60, 327–346.

Goldstein, H. (1983). Measuring changes in educational attainment over time: problems and possibilities. *J. Educ. Meas.* 20, 369–377. doi: 10.1111/j.1745-3984.1983.tb00214.x

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Jpn. Psychol. Res.* 22, 144–149. doi: 10.4992/psycholres1954.22.144

Hanson, B. A., and Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Appl. Psychol. Meas.* 26, 3–24. doi: 10.1177/0146621602026001001

He, W., and Reckase, M. D. (2014). Item pool design for an operational variable-length computerized adaptive test. *Educ. Psychol. Meas.* 74, 473–494. doi: 10.1177/0013164413509629

Hu, H., Rogers, W. T., and Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Appl. Psychol. Meas.* 32, 311–333. doi: 10.1177/0146621606292215

Kaskowitz, G. S., and de Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Appl. Psychol. Meas.* 25, 39–52. doi: 10.1177/01466216010251003

Kim, S. H., and Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *J. Educ. Meas.* 29, 51–66. doi: 10.1111/j.1745-3984.1992.tb00367.x

Kolen, M. J., and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*, 3rd Edn. New York, NY: Springer, doi: 10.1007/978-1-4939-0317-7_10

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Loyd, B. H., and Hoover, H. D. (1980). Vertical equating using the rasch model. *J. Educ. Meas.* 17, 179–193. doi: 10.1111/j.1745-3984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *J. Educ. Meas.* 14, 139–160. doi: 10.1111/j.1745-3984.1977.tb00033.x

Miller, G. E., and Fitzpatrick, S. J. (2009). Expected equating error resulting from incorrect handling of item parameter drift among the common items. *Educ. Psychol. Meas.* 69, 357–368. doi: 10.1177/0013164408322033

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Econ. Rev.* 51, 1–23.

Ogasawara, H. (2002). Stable response functions with unstable item parameter estimates. *Appl. Psychol. Meas.* 26, 239–254. doi: 10.1177/0146621602026003001

R Core Team (2018). *R: A Language and Environment for Statistical Computing [Software]*. Vienna: R Foundation for Statistical Computing.

Revuelta, J., and Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *J. Educ. Meas.* 35, 311–327. doi: 10.1111/j.1745-3984.1998.tb00541.x

Segall, D. O. (2005). "Computerized adaptive testing," in *Encyclopedia of Social Measurement*, ed. K. Kempf-Leonard (Boston: Elsevier Academic), 429–438. doi: 10.1016/b0-12-369398-5/00444-8

Stocking, M. L., and Lewis, C. L. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *J. Educ. Behav. Stat.* 23, 57–75. doi: 10.3102/10769986023001057

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. *Appl. Psychol. Meas.* 7, 201–210. doi: 10.1177/014662168300700208

Sympson, J. B., and Hetter, R. D. (1985). "Controlling item exposure rates in computerized adaptive testing," in *Proceedings of the 27th Annual Meeting of the Military Testing Association*, (San Diego, CA: Navy Personnel Research and Development Center), 973–977.

Thissen, D., Steinberg, L., and Wainer, H. (1988). "Use of item response theory in the study of group difference in trace lines," in *Test Validity*, eds H. Wainer and H. Braun (Hillsdale, NJ: Lawrence Erlbaum Associates).

Thompson, N. A., and Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Pract. Assess. Res. Eval.* 16:9.

Vale, C. D., Maurelli, V. A., Gialluca, K. A., Weiss, D. J., and Ree, M. J. (1981). *Methods for Linking Item Parameters (AFHRL-TR-81-10)*. Brooks Air Force Base TX: Air Force Human Resources Laboratory.

van der Linden, W. J. (2016). *Handbook of Item Response Theory*, Vol. 1. London: Chapman and Hall.

van der Linden, W. J., and Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Appl. Psychol. Meas.* 22, 259–270. doi: 10.1177/01466216980223006

Weeks, J. P. (2010). plink: an r package for linking mixed-format tests using IRT-based methods. *J. Stat. Softw.* 35, 1–33. doi: 10.18637/jss.v035.i12

Wingersky, M. S., and Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Appl. Psychol. Meas.* 8, 347–364. doi: 10.1177/014662168400800312

Yousfi, S., and Böhme, H. F. (2012). Principles and procedures of considering item sequence effects in the development of calibrated item pools: conceptual analysis and empirical illustration. *Psychol. Test Assess. Model.* 54, 366–393.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor declared a shared affiliation, though no other collaboration, with one of the authors AFr at the time of review.

Copyright © 2019 Born, Fink, Spoden and Frey. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Anhang C: Beitrag 3 – Determinants of higher education teachers' intention to use technology-based exams

Zitation: Fink, A., Spoden, C. & Frey, A. (2022). Determinants of higher education teachers' intention to use technology-based exams. *International Journal of Educational Technology in Higher Education.* Manuskript eingereicht zur Publikation am 03.05.2022.

Determinants of Higher Education Teachers' Intention to Use Technology-Based Exams

Aron Fink[1], Christian Spoden[2], and Andreas Frey[1]

[1]Goethe University Frankfurt, Frankfurt am Main, Germany

[2] University of Applied Science Emden/Leer, Emden, Germany

Author Note

Aron Fink, Educational Psychology with focus on Counselling, Measurement, and Evaluation, Institute of Psychology, Goethe University Frankfurt, Germany; Christian Spoden, Psychology with focus on General Psychology, Psychological Methods and Differential Psychology, Department of Business Studies, University of Applied Science Emden/Leer, Emden, Germany; Andreas Frey, Educational Psychology with focus on Counselling, Measurement, and Evaluation, Institute of Psychology, Goethe University Frankfurt.

Correspondence concerning this article should be addressed to Aron Fink, Institute of Psychology, Department of Educational Psychology, Theodor-W.-Adorno-Platz 6, 60629 Frankfurt, Germany, Email: a.fink@psych.uni-frankfurt.de

Abstract

The replacement of existing technology or the introduction of novel technology into the day-to-day routines of higher education institutions is not a trivial task. Currently, many higher education institutions are faced with the challenge of replacing existing procedures for administering written exams with e-exams. To guide this process, this paper proposes the novel technology-based exams acceptance model (TEAM) and empirically evaluates its model structure and usefulness from the perspective of higher education teachers. The model can be used to guide the transition from paper-based exams to e-exams and the implementation of innovative (e.g., adaptive) e-exam formats. The model includes perceived usefulness, computer self-efficacy, computer anxiety, prior experience, facilitating conditions, and subjective norm as predictors of the behavioral intention to use e-exams. To test the model empirically, the responses of 992 teachers at 63 German universities to a standardized online questionnaire were analyzed using structural equation modeling. The model fit was acceptable. With 78% (conventional e-exams) and 92% (adaptive e-exams), a large proportion of the variance of the intention to use these types of exams was explained. With TEAM, a highly predictive model for explaining the behavioral intention to use e-exams is now available. It offers a theoretical basis that can be used for the successful implementation of e-exams in higher education.

Keywords: Technology-based assessment, Adaptive testing, Higher education, Performance-based assessment, Educational technology

Determinants of Higher Education Teachers' Intention to Use Technology-Based Exams

Introduction

Digital transformation affects nearly all areas of modern life. This also includes higher education. When it comes to testing student performance in terms of written examinations, a shift from paper-based exams to computer-administered exams has begun. Furthermore, the COVID-19 pandemic and the associated forced shift to digital learning in higher education institutions around the globe can be seen as a tipping point for the integration of digital technologies for exam purposes into higher education (St-Onge et al., 2021).

The use of computer-based assessments (CBA) as summative assessment tools in higher education can be subsumed under the term e-exams. In this article, e-exams are defined as timed, summative high-stakes assessments of student performance that use digital devices to run a standardized assessment system and in which responses are submitted and, for the most part, scored electronically (e.g., Fluck, 2019). The use of e-exams has several advantages compared to conventional paper-based exams, such as increased test security, cost and time reduction, automated test assembly and analysis of test responses, rapid or even immediate feedback on results, the possibility to integrate interactive elements and multimedia tools into the assessment process, provision of more authentic tests, and the possibility of automatic record keeping for item analysis (Boevé et al., 2015; Nikou & Economides, 2018a; Rolim & Isaias, 2019; Spoden & Frey, 2021; St-Onge et al., 2021). Furthermore, compared to conventional paper-based testing, using CBA can have positive effects on students' test-taking motivation, self-efficacy, test perception, and even test performance (Chua & Don, 2013; Gu et al., 2020; Nardi & Ranieri, 2018; Nikou & Economides, 2016; Rolim & Isaias, 2019).

In addition, the use of digital technologies for examination purposes provides the opportunity to integrate state-of-the-art methods from psychometrics and psychological evaluation into the testing practice (see, e.g., Spoden & Frey, 2021). In particular, new

methods for item calibration such as the continuous calibration strategy (Born et al., 2019; Fink et al., 2018; Frey & Fink, in press) make it possible to combine conventional e-exams with the modern assessment technology of computerized adaptive testing (CAT; Frey, in press). CAT is a testing mode in which the selection of the items to be presented to the test taker depends on the test taker's responses to the most recent items administered. Therefore, the tests are tailored to the individual ability level of the test takers. This typically yields substantially higher measurement precision and/or a shorter test length compared to conventional nonadaptive testing (Segall, 2005). Furthermore, conventional nonadaptive tests typically have the problem that they provide the highest measurement precision for test takers of medium ability, while the precision decreases for test takers with high or low test scores (Dolan & Burling, 2017); adaptive e-exams can help to solve this problem by aligning the standard errors across the complete ability range. Thus, adaptive e-exams can provide teachers in higher education with highly reliable measures of student abilities.

However, the implementation of e-exams that leverage current assessment advancements and digital technologies in higher education is not trivial. Rather, several conditions have to be met for this to happen. Research from the area of technology acceptance (e.g., Marangunić & Granić, 2015) provides reference points for the necessary conditions for successful technology integration in general (e.g., Abdullah & Ward, 2016; Al-Emran et al, 2018; Granić & Marangunić, 2019; Scherer & Teo, 2019). Technology acceptance comprises different attitudes regarding technology and beliefs that explain a person's intentions to use technology, as well as their actual use of technology (Davis, 1989). One model in particular has dominated the research on factors that influence the acceptance and use of technologies: the technology acceptance model (TAM; Davis, 1989). The core assumption of TAM is that perceived usefulness (PU) and perceived ease of use (PEOU) are the central factors that influence a person's attitude toward and behavioral intentions with regard to technology use.

However, TAM and related models focus on technology use in general and not

specifically on e-exams. Therefore, in their current form, they do not cover the circumstances relevant for the implementation of e-exams precisely enough. A specific TAM for e-exams in higher education would be very useful for many higher education institutions, where decisions have to be made that make the successful implementation of e-exams possible. The successful implementation of e-exams in the routine procedures of higher education institutions depends, among others, on the technology acceptance of the stakeholders involved. If they do not accept this assessment type, they are often in positions in which they can prevent its implementation regardless of the advantages.

As previous research on the implementation of e-exams in higher education has primarily focused on the acceptance of e-exams by students (Maqableh et al., 2015; Terzis & Economides, 2011; Terzis et al., 2012; Zheng & Bender, 2019), there is a need for studies that explicitly consider the perspectives of the academic staff on e-exams (Bennett et al., 2017; Brady et al., 2019; Deeley, 2018). Especially the viewpoint of the teaching staff, as they are responsible for the design and integration of e-exams into courses, is a critical factor for the successful implementation of e-exams (Bennett et al., 2017; Brady et al., 2019; Nikou & Economides, 2018a; Paiva et al., 2017). In order to avoid time-consuming and expensive failures during the implementation process, a thorough understanding of the conditions necessary for teachers in higher education to accept e-exams as viable evaluation tools and, therefore, to form a strong intention to use them is exceptionally important. In addition, among the few studies that examined the perspectives of higher education teaching staff on e-exams, even fewer are situated in a clearly defined theoretical framework (Brady et al., 2019).

Against this background, this study aimed to formulate a specific theoretical model on the acceptance of e-exams that makes it possible to predict the behavioral intention of higher education teachers to use e-exams. This model is called the technology-based exams acceptance model (TEAM). It draws from TAM and its extensions (e.g., Terzis & Economides, 2011; Venkatesh et al., 2003). TEAM is intended for use in guiding

implementation processes to make them successful. In order to justify such use, empirical data was gathered and statistically analyzed with structural equation modeling to test whether the proposed model structure fits the actual response behavior of higher education teachers. After establishing the model structure, TEAM was used to examine the as yet unanswered question of whether different conditions have to be met before implementing innovative adaptive e-exams compared to the conditions that need to be met before implementing conventional e-exams, which basically mimic paper-pencil exams with computers.

The study had the following four research objectives (ROs):

RO1: To formulate TEAM.

RO2: To examine the appropriateness of TEAM for teachers in higher education.

RO3: To statistically test the theoretically derived direct and indirect effects described by TEAM.

RO4: To examine whether there are differences with regard to the structure and the path coefficients of TEAM between adaptive e-exams and conventional e-exams.

The text is organized as follows: The next section describes theoretical perspectives on technology acceptance models and the most relevant previous studies on technology acceptance in education. Based on this literature review, hypotheses are derived and TEAM is formulated. The following section covers the methods used to test the hypotheses and to examine the model. Subsequently, the results are presented. Finally, the results are discussed regarding the research objectives, along with practical implications and pathways for future research on TEAM.

Educational Technology Acceptance Model

TAM is the most frequently used theory in technology acceptance literature in general (e.g., Marangunić & Granić, 2015) and in e-learning acceptance literature in particular (e.g., Abdullah & Ward, 2016; Granić & Marangunić, 2019; Scherer et al., 2019). In the context of

educational technologies, numerous studies have explored the applicability of TAM and connected models across a broad range of technologies. Among these are, for instance, mobile learning (Sánchez-Prieto et al., 2016; Mutambara & Bayaga, 2021), digital learning environments (Bauwens et al., 2020; del Barrio-García et al., 2015), learning management systems (Alharbi & Drew, 2014; Cigdem & Topcu, 2015; Sánchez & Hueros, 2010), multimedia platforms adapted for learning (Lee & Lehto, 2013), communication and collaboration applications (Maican et al., 2019), virtual reality (Noble et al., 2022), as well as CBA (Maqableh et al., 2015; Terzies, & Economides, 2011) or mobile-based assessment (Nikou & Economides, 2017; Nikou & Economides, 2018b).

TAM originates in the theory of reasoned action (Ajzen & Fishbein, 1980). It comprises several variables that directly or indirectly explain the behavioral intention to use technology and the actual use of technology. In the original model, Davis (1989) suggested that three factors influence technology use: perceived ease of use (PEOU), perceived usefulness (PU), and attitude toward using (ATU). PU is defined as a person's belief about the degree to which using the particular system would enhance their job performance. PEOU is a person's belief about the degree to which using the particular system would be free of effort (Davis, 1989). Davis (1989) hypothesized that ATU is the main determinant of technology use. PU and PEOU are considered to influence ATU.

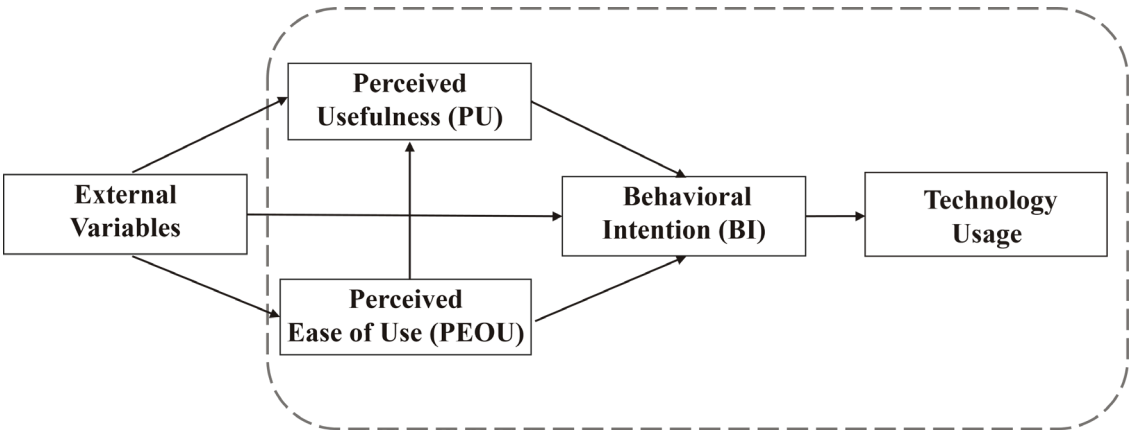
Subsequent TAM developments added the behavioral intention (BI) of a person to use a particular system as a new variable that was directly influenced by the PU of the system (Davis et al., 1989). In addition, Davis et al. (1989) argued that there would be cases where an individual could form a strong BI to use the particular system that was perceived as useful without forming any kind of attitude toward using the system and, thus, he removed the ATU construct from the model. In line with this, a large body of studies has underlined the weak role of ATU as a mediator between BI, PU, and PEOU (Davis et al., 1989; Szajna, 1996; Teo, 2009; Venkatesh & Davis, 2000; Wang & Wang, 2009; Yen et al., 2010). Therefore,

Venkatesh (2000) considered the simplified version of TAM to be superior to the original model in predicting user acceptance, by including the direct effects of both PU and PEOU on BI.

TAM has been modified and extended by different authors over the last decades by having external factors added that explain variation in the TAM core variables PU, PEOU, or BI (Figure 1; see Marangunić & Granić, 2015 for an overview). This has resulted in a large number of different external factors and several extended TAMs in the research area of e-learning acceptance (Abdullah & Ward, 2016). These external factors represent individual characteristics and beliefs as well as contextual factors. Among others, subjective norm, facilitating conditions, computer self-efficacy, computer anxiety, and prior experience are the most commonly used external factors in the context of e-learning that have been found to be significantly related to the TAM core variables by means of meta-analyses (Abdullah & Ward, 2016; Schepers & Wetzels, 2007; Scherer et al., 2019) or in the systematic review by Granić and Marangunić (2019). These theoretical considerations and empirical findings are presented and discussed in the next section to formulate an empirically testable model of possible factors that influence the intention to use e-exams, both conventional and adaptive, from the perspective of higher education teachers (Figure 2).

Figure 1

Technology Acceptance Model (TAM)



Hypotheses and Model Derivation

Perceived Usefulness (PU)

PU is one of the core variables in TAM and its extensions. There is solid empirical evidence for a positive effect of PU on teachers' BI to use educational technologies (e.g., Cigdem & Topcu, 2015; Harris et al., 2016; Lin et al., 2013; Motaghian et al., 2013). In addition, Granić and Marangunić (2019) identified PU as the strongest determinant for the adoption of various learning technologies in their systematic review. We expected to find a similar effect for e-exams. Therefore, the first hypothesis was:

Hypothesis 1: PU has a positive effect on BI.

Prior Experience (PE)

King and He (2006) stated that prior experience is one of the best-studied external factors in the TAM context. Research indicates that individuals with more computer-related experience, such as those who use a computer to write emails or who use word processing software, spreadsheet programs, and others, are more likely to show a higher degree of PU and PEOU with regard to a new e-learning system (Abdullah & Ward, 2016; Lee et al., 2013). As it can be assumed that teachers in higher education often carry out such computer activities in their everyday work, prior experience was more explicitly specified for the higher education setting to comprise the prior use of multimedia and technology for teaching purposes instead of computer-related experience in general. We hypothesized that:

Hypothesis 2: PE has a positive effect on PU.

Computer Self-Efficacy (CSE)

CSE is defined as a person's belief about the ease with which they ca perform a specific task using a computer (Compeau & Higgins, 1995). CSE can affect the BI to use computers because people who believe that they do not have the ability to use computers will avoid using them (Igbaria & Iivari, 1995; Kwon et al., 2007). Conversely, the higher a person's CSE is, the higher that person's use of computers will be (Compeau & Higgins,

1995). Ahmad et al. (2010) showed that this connection also holds with regard to the adoption of e-learning by teachers. In addition, the meta-analysis of teachers' adoption of e-learning conducted by Scherer et al. (2019) identified CSE as one of the strongest antecedents of PU and PEOU. Conceptually, PEOU and CSE have a lot of similarities. As stated above, PEOU reflects the degree to which a person believes that the system of interest is easy to use, which, in turn, is also reflected by their CSE. These commonalities have been supported by empirical studies, and both constructs are sometimes even measured with similar items (Scherer et al., 2015; Scherer & Teo, 2019). In particular, if a study is not about an existing e-exam system but about the hypothetical use of such a system, PEOU can be regarded as an expression of CSE rather than as the actual ease of use of the corresponding system. Therefore, we integrated CSE rather than PEOU into the model and investigated its impact on PU and BI. In addition, as low CSE is assumed to lead to a lower degree of computer use in general, it can be assumed that teachers in higher education with low CSE use less technology for teaching purposes in general and, therefore, show less PE. We hypothesized that:

Hypothesis 3: CSE has a positive effect on PE.

Hypothesis 4: CSE has a positive effect on PU.

Hypothesis 5: CSE has a positive effect on BI.

Computer Anxiety (CA)

CA is defined as the degree of apprehension or even fear that an individual experiences when using a computer (Venkatesh & Morris, 2000). In this study, CA is regarded as a time-persistent trait that contains both cognitive and affective components (e.g., Morris et al., 1981; Richter et al., 2010). A number of studies have shown that CA is associated with the avoidance or reduced use of e-learning systems (Abdullah & Ward, 2016). A discrepancy between educators' perceptions of their technological competence and the learning effort they have to put into using computers for teaching purposes can often be perceived as threatening and overwhelming. Thus, the anxiety of a teacher in higher education

affects both the extent to which and the way in which they use technology in everyday instructional practice (Al-alak & Alnawas, 2011; Mac Callum et al., 2014). Therefore, it can be assumed that teachers with high CA gather less PE. In addition, the lack of technology use due to CA can be assumed to prevent the development of a high degree of CSE (Lee & Huang, 2014). The next two hypotheses were therefore:

Hypothesis 6: CA has a negative effect on CSE.

Hypothesis 7: CA has a negative effect on PE.

Subjective Norm (SN)

SN is defined as a person's perception that most people who are important to them think that they should show the behavior in question (Fishbein & Ajzen, 1975). With regard to e-exam adoption by higher education teachers, SN can be regarded as the extent to which a higher education teacher perceives pressure from members in their environment (e.g., colleagues, students, or the administrative staff) to use e-exam systems. The perception of such pressure increases the likelihood to incorporate positive beliefs regarding an e-exam system into one's own beliefs system. It also increases the probability to perceive the system as useful and to form a strong BI to use it. Prior research on higher education teachers' e-learning adoption supports this assumption and identified SN as an important determinant of PU and BI (Cigdem & Topcu, 2015; Garcia & Gomez, 2014; Nikou & Economides, 2018b; McGill et al., 2011; Motaghian et al., 2013; Wang & Wang, 2009). We hypothesized that:

Hypothesis 8: SN has a positive effect on PU.

Hypothesis 9: SN has a positive effect on BI.

Facilitating Conditions (FC)

FC is defined as a person's perception of the degree to which organizational and technical resources exist to support the use of a particular technology. Depending on the system, FC comprises many different aspects and is typically operationalized to include aspects of the environment that are designed to remove barriers to using the technology

(Venkatesh et al., 2003). The aspects that are relevant for this study are, especially, the provision of organizational and technical support (e.g., skills training, information and supportive material, administrative support, availability of a designated person to help, etc.) and appropriate technical resources for carrying out e-exams (e.g., hardware, software, intranet). If any of these elements are perceived as missing, a person can avoid forming the intention to use an e-exam system. Conversely, it can be assumed that the more supportive the existing conditions are, the more likely it is that a higher education teacher will intend to use an e-exam system. In line with this, Lin et al. (2013), for example, found FC to have a positive effect on higher education teachers' BI to use podcasting for e-learning. Thus, we hypothesized:

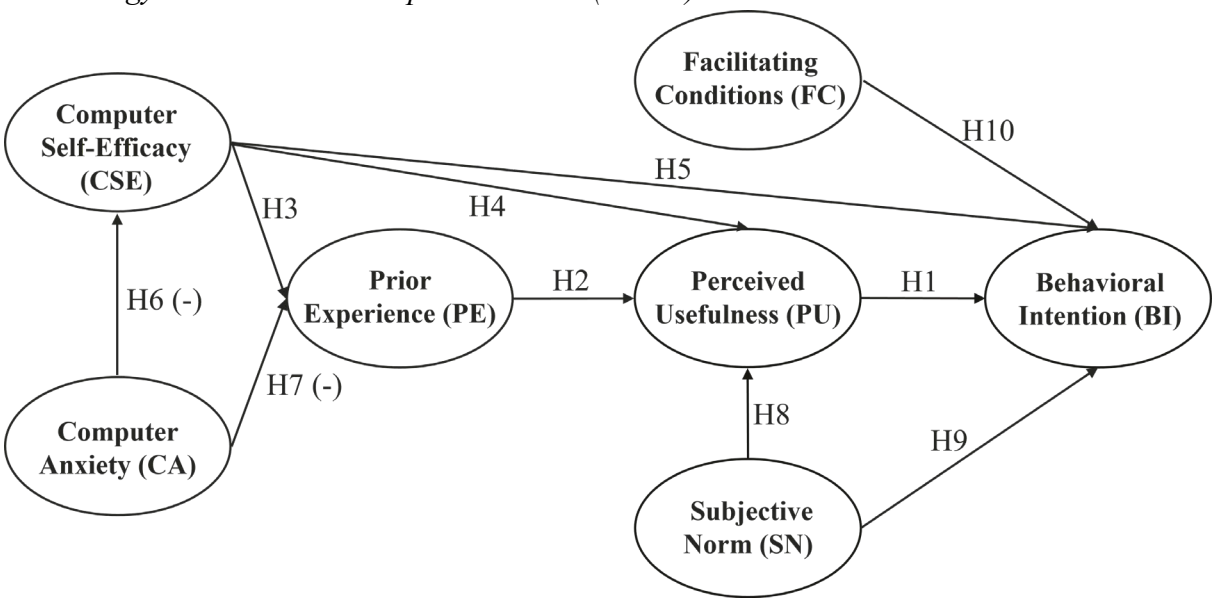
Hypothesis 10: FC has a positive effect on the BI.

The complete research model including these hypotheses is presented in Figure 2.

Because the model is intended to be applicable to both the transition from paper-based exams to e-exams and the implementation of innovative adaptive e-exams, it was further assumed that the effects would be the same for conventional and adaptive e-exams. Thus, the last hypothesis was:

Figure 2

Technology-Based Exams Acceptance Model (TEAM)



Note. (-) = negative effect. H1–H10 = Hypothesis 1–10.

Hypothesis 11: The above-mentioned effects are invariant between conventional and adaptive e-exams.

Materials and Methods

Sample

To test the formulated hypotheses, a nationwide (Germany), cross-disciplinary, and cross-institution online study, which addressed the staff involved in teaching, was conducted. The sample was acquired via email by contacting the secretaries of every institute of at least two universities and two universities of applied science per federal state and requesting them to forward the questionnaire invitation to the teaching staff of the institute. Participation was voluntary. The overall sample comprised $N = 1{,}000$ higher education teachers but eight participants were excluded from the analysis due to missing data on almost all items. Thus, the final sample comprised $N = 992$ (sex: 38 % female; age: $M = 44.29$, $SD = 11.87$) higher education teachers distributed across all 16 federal states of Germany, 63 higher education institutions, and 35 disciplines. The participants were randomly assigned to one of two groups. Group 1 comprised $N_{ex} = 494$ higher education teachers who responded to a questionnaire on conventional e-exams and Group 2 comprised $N_{ad} = 498$ higher education teachers who responded to a questionnaire on adaptive e-exams (see Table 1 for demographic information).

Measures

Along with some demographic questions, both online questionnaires comprised multiple-item scales, ensuring a reliable measurement of the above-mentioned constructs. A mixture of existing scales and scales developed anew by the authors was used. For CSE and PE, corresponding scales from the German version of the teacher questionnaire used in the International Computer and Information Literacy Study 2013 (ICILS; Gerick et al., 2018) were used. As the original items were developed to be used on secondary school teachers, the wording of the items was adapted to the context of higher education. The changes made in the

Table 1
Demographics of the Two Subsamples

	Group	
Variable	Group 1 (e-exams)	Group 2 (adaptive e-exams)
N	494	498
Sex	female: 38.9 %	female: 37.2 %
Average age (*SD*)	44.30 (11.95)	44.29 (11.80)
Highest academic degree		
• Post-doctoral degree	23.3 %	22.9 %
• Doctoral degree	44.9 %	47.2 %
• Master's degree	30.2 %	27.7 %
• Bachelor's degree	0.4 %	1.0 %
• Other	1.2 %	1.2 %
Already used e-exams for summative assessments	15.6 %	15.1 %

German version of the items are given in Appendix A (Table A.1). For CSE, the participants had to rate different computer-related tasks on a three-point Likert scale with the response categories *I know how to do this*, *I could work out how to do this*, and *I do not think I could do this.* For PE, the participants were asked to indicate how often they used different information and communication technology (ICT) tools when teaching on a four-point Likert scale with the response categories *never*, *in some lessons*, *in most lessons*, and *in every or almost every lesson*. The instrument for CA was adapted from Richter et al. (INCOBI-R; 2010). The items had a five-point Likert scale ranging from *strongly disagree* to *strongly agree* with the middle category labeled *neutral*. The remaining scales for PU, SN, FC, and BI were developed by the authors. Prior to their use in the study presented here, these scales were trialed and optimized (N_{pre} = 109 teachers from a German university; Klösel, 2018). The items of these scales had a four-point Likert scale, ranging from *totally disagree* to *totally agree*. The respondents had to provide an answer to every item. The item wording of the developed scales with their English

translation is given in Appendix A (Table A.2). As it could not be assumed that each of the participants had a deeper understanding of the concepts of e-exams and adaptive e-exams, both terms were explained in the online questionnaires and the main advantages and disadvantages were mentioned.

Procedure

At the beginning of the study, the participants were provided with information about the study and were asked for consent for their data to be used in the study. Afterwards, each participant was asked to answer an online questionnaire. Two versions of the online questionnaires were used; the first one focused on conventional e-exams and the second one on adaptive e-exams. The online questionnaire versions were assigned randomly to the participants. As mentioned above, the participants were required to complete all items in the questionnaire and were not allowed to skip items. They generally completed the questionnaire within 20 minutes.

Data Analysis

Once the data were gathered, the explanatory model was tested by means of multigroup structural equation modeling (MG-SEM) in Mplus 8.1 (Muthén & Muthén, 2018). Weighted least square mean and variance adjusted (WLSMV) estimation was used to model the ordinal data. We used the Mplus TYPE = COMPLEX procedure, which adjusts model fit statistics and standard errors for error dependencies caused by the clustered structure of the data. The clusters in our data set were the universities, with the higher education teachers nested in them. As participants were required to complete all items except the items asking for demographic information, only a few items had missing responses (due to test aborts) and, in those cases, the missing rates were very low (< 0.5% per item). The few missing responses were assumed to be missing completely at random and were treated by pairwise deletion as implemented in Mplus when using WLSMV. As a first step, we estimated the measurement model (simple structure with correlated factors, see Appendix B) in both groups (Group 1 [e-

exams], Group 2 [adaptive e-exams]) via multigroup confirmatory factor analysis (MG-CFA), and we examined the loading pattern. Additionally, when testing the hypotheses and comparing the groups, we conducted a measurement invariance analysis to examine whether the constructs had been measured in a directly comparable manner in the two groups. In a second step, we conducted the full MG-SEM and inspected model fit (RO1). In this analysis, we estimated the model shown in Figure 2. Afterwards, the statistical significance of the proposed relations was examined separately for each group (Hypotheses 1–10). Hypothesis 11 was tested by means of Wald tests that compared latent means and path coefficients between the two groups.

Results

Descriptive Results and Measurement Model

Table 2 shows the descriptive results as well as Green and Yang's (2009) variation of coefficient ω as a measure of reliability for categorical data. The ω exceeded the suggested rule of thumb of .70 for all scales, so that the reliability of all scales can be regarded as acceptable or better. The standardized factor loadings of the items ranged from .631 to .943 for e-exams and from .617 to .953 for adaptive e-exams (see Appendix B for a detailed presentation of the results of the measurement model). The fit of the measurement model (see fit measures of the configural model in Table 3) can be regarded as acceptable.

Measurement Invariance Analysis

In order to test the measurement invariance of the proposed measurement model across the two groups, we used the four-step approach to test measurement invariance (e.g., van de Schoot et al., 2012). This includes analyzing (1) configural invariance (noninvariance model), (2) metric invariance (invariant factor loadings across groups), (3) scalar invariance (invariant factor loadings and thresholds across groups), and (4) residual invariance (invariant factor loadings, thresholds, and residual variances across groups). Measurement invariance is

Table 2

Means, Standard Deviations, Reliability, and Latent Correlations

Group	Factor	*M*	*SD*	ω	BI	PU	PE	CSE	CA	SN	FC
Group 1 (e-exams)	BI	2.41	0.82	.881	1.00	-	-	-	-	-	-
	PU	2.36	0.72	.874	**.867**	1.00	-	-	-	-	-
	PE	1.67	0.46	.858	**.271**	**.255**	1.00	-	-	-	-
	CSE[a]	2.65	0.32	.853	**.197**	**.217**	**.460**	1.00	-	-	-
	CA	1.50	0.52	.871	-.014	-.041	**-.144**	**-.497**	1.00	-	-
	SN	2.11	0.48	.717	**.602**	**.605**	**.287**	**.241**	.035	1.00	-
	FC	2.00	0.72	.873	**.331**	**.275**	**.199**	**.232**	-.007	**.652**	1.00
Group 2 (adaptive e-exams)	BI	2.22	0.77	.916	1.00	-	-	-	-	-	-
	PU	2.05	0.76	.940	**.852**	1.00	-	-	-	-	-
	PE	1.68	0.45	.809	**.219**	**.268**	1.00	-	-	-	-
	CSE[a]	2.68	0.32	.835	**.155**	**.155**	**.463**	1.00	-	-	-
	CA	1.48	0.50	.832	**.097**	.087	**-.112**	**-.513**	1.00	-	-
	SN	2.19	0.53	.722	**.244**	**.338**	**.219**	.079	.052	1.00	-
	FC	2.12	0.82	.902	**.098**	**.120**	**.212**	.073	.002	**.620**	1.00

Note. Correlation coefficients significantly different from 0 ($p \leq .05$) are printed in bold. ω = Green and Yang's (2009) variation of coefficient ω; BI = Behavioral intention to use; PU = Perceived usefulness, PE = Prior experience; CSE = Computer self-efficacy; CA = Computer anxiety; SN = Subjective norm; FC = Facilitating conditions.

[a] Six items were removed due to limited variance.

usually determined by testing whether the difference in the global model fit between the compared groups, $\Delta\chi^2$, differs from zero to a statistically significant extent (Byrne et al., 1989). However, because $\Delta\chi^2$ is sensitive to sample size, Chen (2007) recommends using the change in alternative globalmodel fit indices as a criterion as well. Chen suggests a criterion of a -.01 maximum change in the comparative fit index (CFI), together with changes in the root mean squared error of approximation (RMSEA) .015 and the standardized root mean square residual (SRMR) of .030, for metric invariance or .015 for scalar or residual invariance.

Table 3 shows the results of the measurement invariance tests: the residual invariance of the factors across the two groups can be regarded as established, with all changes in alternative global model fit indices smaller than the criteria mentioned above.

Table 3

Result of Measurement Invariance Tests for the Two Analyzed Groups

Model	χ^2 (*df*)	CFI	RMSEA (90% CI)	SRMR	$\Delta\chi^2$ (Δdf)	ΔCFI	ΔRMSEA	ΔSRMR
configural	4788.331 (3154)	.948	.032 (.030/.034)	.090	-	-	-	-
metric	4903.141 (3205)	.947	.032 (.030/.034)	.090	114.810* (51)	-.001	.000	.000
scalar	5118.789 (3366)	.946	.032 (.031/.034)	.091	215.648* (161)	-.001	.000	.001
residual	5148.118 (3380)	.943	032 (.031/.034)	.093	29.329* (14)	-.002	.000	.002

Note. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

* $p \leq .05$.

Overall Model Fit

Following the investigation of the measurement model and the measurement invariance analysis, the full MG-SEM was estimated. As a first step, the overall global model fit of the research model was evaluated. The results indicated that the research model had an acceptable fit: $\chi^2 = 5122.318$, $df = 3324$, CFI = .947, TLI = .948, SRMR = .095, RSMEA = .033 (90% CI [.031, .035]). In both groups, a very large proportion of the BI was explained with the suggested

model (Group 1 [e-exams]: 77.5%; Group 2 [adaptive e-exams]: 91.6%). Thus, TEAM can be regarded as robust and as being able to explain higher education teachers' intention to use (adaptive) e-exams well.

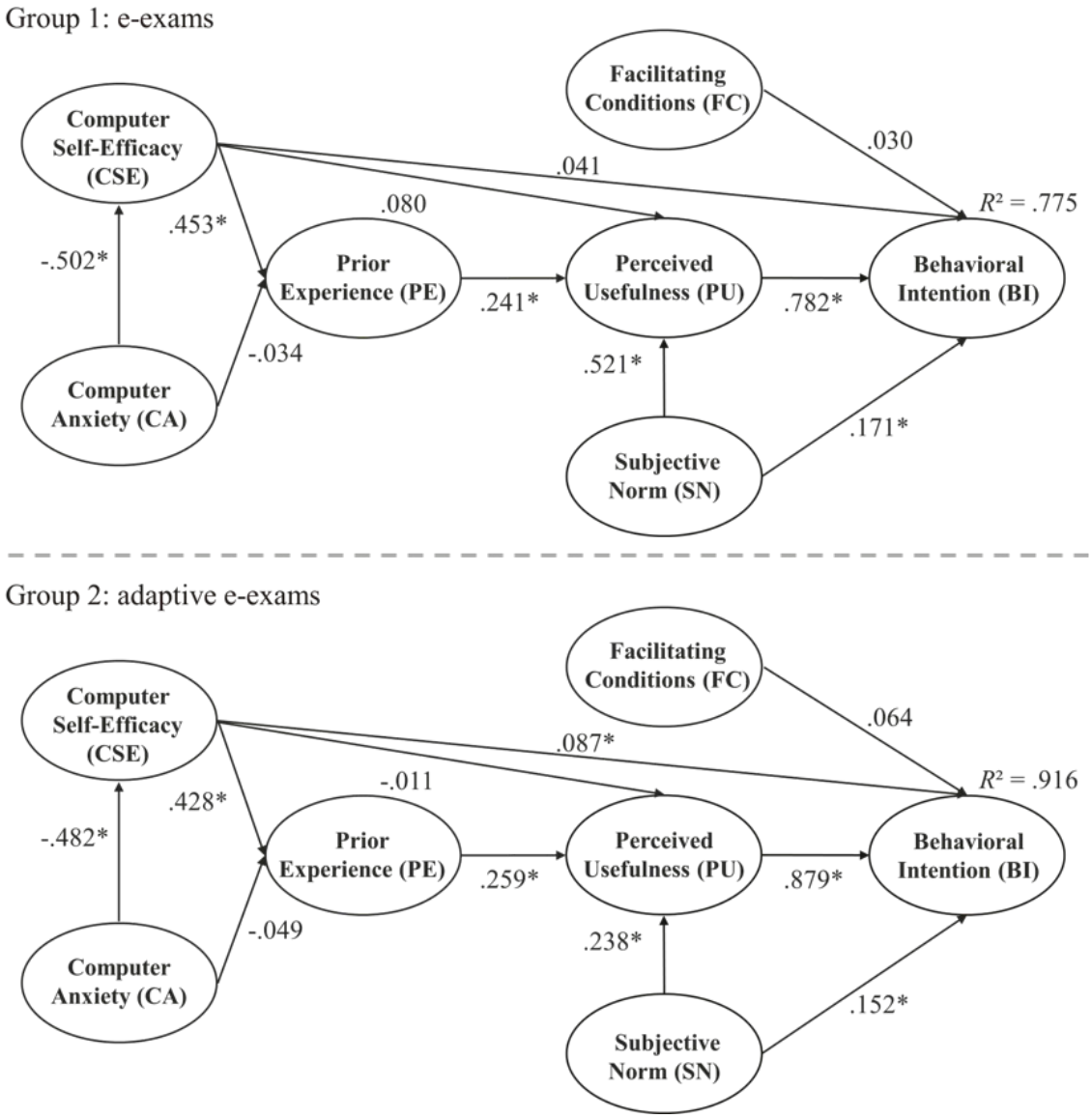
Hypotheses Testing

As a second step, and addressing RO2, the proposed relations were statistically tested separately for each group. Figure 3 summarizes the results for the hypotheses. In both groups, PU had a strong positive effect on BI (Hypothesis 1; Group 1 [e-exams]: $p < .001$; Group 2 [adaptive e-exams]: $p < .001$). PE had a positive effect on PU for conventional e-exams as well as for adaptive e-exams (Hypothesis 2; Group 1 [e-exams]: $p < .001$; Group 2 [adaptive

e-exams]: $p < .001$). Regarding CSE, as hypothesized, in both groups, a direct positive effect on PE (Hypothesis 3; Group 1 [e-exams]: $p < .001$; Group 2 [adaptive e- exams]: $p < .001$) was found but there were no direct effects on PU (Hypothesis 4; Group 1 [e-exams]: $p = .220$; Group 2 [adaptive e-exams]: $p = .843$). In the adaptive e-exams group, CSE had a small positive effect on BI, which was not the case for conventional e-exams (Hypothesis 5; Group 1 [e-exams]: $p = .204$; Group 2 [adaptive e-exams]: $p = .002$). In addition, in both groups, CSE had a significant indirect effect on PU that was mediated by PE (Group 1 [e-exams]: β_{ind}

Figure 3

Results for the MG-SEM Analysis



Note. * $p \leq .05$, standardized path coefficients.

= .109, $p < .001$; Group 2 [adaptive e-exams]: $\beta_{ind} = .111$, $p < .001$) but it did not have an indirect effect on BI via PU (Group 1 [e-exams]: $\beta_{ind} = .062$, $p = .225$; Group 2 [adaptive e-exams]: $\beta_{ind} = .-.010$, $p = .843$). CA was found to have a medium direct negative effect on CSE in both groups (Hypothesis 6 Group 1 [e-exams]: $p < .001$; Group 2 [adaptive e-exams]: $p < .001$). There was no direct effect of CA on PE (Hypothesis 7; Group 1 [e-exams]: $p = .418$; Group 2 [adaptive e-exams]: $p = .150$), but there was an indirect effect that was mediated by CSE (Group 1 [e- exams]: $\beta_{ind} = -.227$, $p < .001$; Group 2 [adaptive e-exams]: $\beta_{ind} = -.207$, $p < .001$). As hypothesized, SN had a direct effect on PU (Hypothesis 8; Group 1 [e-exams]: $p < .001$; Group 2 [adaptive e-exams]: $p < .001$) as well as on BI (Hypothesis 9; Group 1 [e-exams]: $p < .001$; Group 2 [adaptive e-exams]: $p = .003$). Next to the direct effect, SN had a significant indirect effect on BI that was mediated by PU in both groups (Group 1: [e-exams]: $\beta_{ind} = .407$, $p < .001$; Group 2 [adaptive e-exams]: $\beta_{ind} = .209$, $p < .001$). Finally, FC did not have an effect on BI (Hypothesis 10; Group 1 [e-exams]: $p = .460$; Group 2 [adaptive e-exams]: $p = .329$). In total, six out of the 10 hypotheses were supported by the results. A summary of the hypotheses testing can be found in Appendix C.

The results clearly identified PU as the key predictor of BI. In addition, SN had small to medium effects on BI and PU. The effect of FC on BI was negligible. Furthermore, PU was directly affected by PE. CSE was found to have only an indirect effect on PU, which was mediated by PE. CSE, in turn, was negatively related to CA.

Comparison of Conventional and Adaptive E-Exams

Comparing the two models, and addressing RO3, our results showed significant standardized latent mean differences between the two factors PU ($d = -.467$, $p < .001$) and BI ($d = -.274$, $p < .001$), with higher values obtained for Group 1 (e-exams) than for Group 2 (adaptive e-exams). There were no significant differences in the other latent means between the two groups. Looking at the path coefficients, the results showed that in Group 1 (e-exams), SN had a significantly larger effect on PU ($\Delta\beta = 0.283$; $p = .007$) and on BI ($\Delta\beta =$

0.019; $p = .008$) than in Group 2 (adaptive e-exams). Apart from this, PU had a significantly larger effect on BI ($\Delta\beta = 0.097$; $p < .001$) in Group 1 (e-exams). The remaining standardized path coefficients did not differ significantly between the two groups. In addition, the intention to use adaptive e-exams was more strongly predicted ($R^2 = 91.6\%$) than the intention to use conventional e-exams ($R^2 = 77.5\%$). Therefore, Hypothesis 11 was only partially supported by our results.

Discussion

Although e-exams have several advantages compared to conventional paper-based exams, problems in the implementation of e-exams can easily become time-consuming and costly for both higher education institutions and teachers. In the current climate of a fundamental shift towards e-exams in more and more higher education institutions, research is urgently needed that provides a profound understanding of the very specific conditions that must be fulfilled to facilitate the implementation of e-exams by teachers in higher education (Bennett et al., 2017; Brady et al., 2019; Deeley, 2018). This is particularly important for the implementation of adaptive e-exams, which offer major advantages in psychometric quality such as substantially improved measurement efficiency, individualization, an extension of the performance bandwidth that can be measured, and an alignment of measurement precision across students. However, the implementation of adaptive e-exams has not yet been covered by previous research specific to higher education. This study makes three contributions to this area. First, with TEAM, it proposes a theoretical model that makes it possible to predict the intention of higher education teachers to use e-exams (RO1). Second, it provides empirical evidence for the appropriateness of the suggested model's structure, including the hypothesized effects and the applicability for both conventional e-exams and adaptive e-exams. Third, it can be applied within change processes at higher education institutions to guide successful implementation processes of e-exams and adaptive e-exams.

Specifically, RO2 aimed to test the general capacity of the proposed model to explain

higher education teachers' intention to use (adaptive) e-exams through its factors by finding evidence in terms of model fit. The model was supported by the data. Thus, TEAM can be regarded as a reliable theoretical basis for explaining higher education teachers' BI to use conventional and adaptive e-exams. With 78% (e-exams) and 92% (adaptive e-exams), large proportions of the variance of BI were explained with the suggested model structure. Therefore, it can be expected that predictions made with TEAM will come very close to the results that can actually be observed. Regarding the influence of the individual factors (RO3), the study resulted in six conclusions:

1. The PU of conventional and adaptive e-exams was the key predictor of the BI to use them.
2. PE in the sense of digital media use in courses led to a higher degree of PU of new educational technologies such as e-exams.
3. CSE did not have a direct effect but did have an indirect effect on PU, which was mediated by PE.
4. High CA led to a lower degree of CSE and, therefore, indirectly and negatively influenced digital media use in courses (PE).
5. The SN resulting from professional social environments played an important role in influencing the BI to use conventional and adaptive e-exams.
6. FC, such as supportive organizational and technical resources, did not have an effect on the BI to use conventional and adaptive e-exams.

RO4 was to investigate whether the results would differ between conventional and adaptive e-exams. The results revealed only a few differences between the two groups. SN had a positive effect on PU in both groups. However, the effect was more than twice as high for conventional e-exams than for adaptive e-exams. The available data do not provide explanations for this. Future studies could investigate this result in more depth, for example, by comparing higher education institutions where adaptive e-exams are used with those where

this is not the case. In addition, the effect of PU on BI was stronger for adaptive e-exams. This could have resulted from the lower effects of SN on BI and PU. Thus, less systematic variance was bound by SN, which could have led to the stronger effect of PU on BI. Moreover, CSE had an effect on the BI to use adaptive e-exams only. Such innovative CBA formats also require the test administrators to have higher technical skills to enable them to follow the principles of CAT. Therefore, it could be assumed that a higher degree of CSE is necessary in order to form a strong BI to use such formats, which results in a positive effect of CSE on BI.

Outlook

The suggested model proved to be capable of explaining the intention of higher education teachers to use (adaptive) e-exams with six interrelated variables. This intention will often directly translate into behavior. However, high BI does not guarantee subsequent behavior, as has been discussed in the TAM-based research (e.g., Liu et al., 2019; Nistor, 2014; Scherer et al., 2020; Wu & Du, 2012). Therefore, future studies should examine whether a strong BI leads to the actual use of e-exam systems. Furthermore, to take into account the complexities of turning intentions into actual behavior, the moderating effects of contextual and social factors on the intention-behavior link should be considered.

TEAM focuses on e-exam acceptance by higher education teachers. The model does not specify which types of professional knowledge about teaching and learning with technology higher education teachers must have in order to integrate technology into the assessment process in a meaningful way. The technological pedagogical content knowledge (TPACK) framework defines the different interrelated knowledge domains necessary for the educationally useful integration of technology into teaching and learning processes (Mishra & Koehler, 2006). Previous studies have shown that TPACK and educational technology acceptance are interrelated (e.g., Hsu, 2016; Mei et al., 2018). Considering this, it would be interesting for future studies to investigate the connections between TPACK and TEAM in

order to get an even deeper understanding of the underlying processes of higher education teachers' e-exam acceptance.

Conclusions

With TEAM, an empirically investigated, highly predictive model for explaining BI is now available. This model offers a sound theoretical basis that can be used to optimize the implementation of e-exams. A promising result of this study is that the higher education teachers in our sample did not express a strong need for expensive infrastructural changes (which would be reflected in stronger effects of FC) in order for them to form a strong BI. Rather, according to the study results, the goals of the implementation process should be a) to promote CSE, for example, by means of academic instruction and training; b) to encourage teachers to try out different kinds of digital media in their courses in order for them to become familiar with them and, thus, to gather experience in using technology for teaching purposes in general; and c) to promote e-exams through an appropriate communication strategy and, therefore, increase the perceived SN. These goals seem achievable as it can be assumed that the COVID-19 pandemic and the related shift to online teaching and learning has forced many countries to vigorously pursue goals a) and b) (e.g., St-Onge et al., 2021), and goal c), in turn, can be supported by already existing structures in higher education institutions. The scales published in this article can also be used to evaluate the effectiveness of interventions that aim to reach these goals.

In conclusion, this study determines the conditions necessary for a successful implementation of e-exams as high-stakes assessments at higher education institutions and it offers the essential building blocks required for a goal-oriented and theory-based implementation of e-exams.

Declarations

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) [grant number 16DHL1005].

Authors' contributions

AFi conceived the study, conducted the acquisition and statistical analyses of the research data, drafted the manuscript, and approved the submitted version. **CS** made substantial contributions to the conceptualization of the study and the interpretation of the study results, reviewed the manuscript critically for important intellectual content, and approved the submitted version. **AFr** made substantial contributions to the conception and design of the study, reviewed the manuscript critically for important intellectual content, and approved the submitted version.

Acknowledgements

Not applicable

References

Abdullah, F., & Ward, R. (2016). Developing a general extended technology acceptance model for e-learning (GETAMEL) by analyzing commonly used external factors. *Computers in Human Behavior, 56,* 238–256. https://dx.doi.org/10.1016/j.chb.2015.11.036

Ahmad, T. B. T., Madarsha, K. B., Zainuddin, A. M., Ismail, N. A. H., & Nordin, M. S. (2010). Faculty's Acceptance of Computer Based Technology: Cross-Validation of an Extended Model. *Australasian Journal of Educational Technology, 26*(2), 268–279. https://doi.org/10.14742/ajet.1095

Ajzen, I., & Fishbein, M. (1980). *Understanding Attitudes and Predicting Social Behavior.* Prentice Hall.

Al-alak, B. A., & Alnawas, I. A. M. (2011). Measuring the acceptance and adoption of E-learning by academic staff. *Knowledge Management & E-Learning: An International Journal, 3*(2), 201–221. https://doi.org/10.34105/j.kmel.2011.03.016

Al-Emran, M., Mezhuyev, V., & Kamaludin, A. (2018). Technology acceptance model in m-learning context: A systematic review. Computers & Education, 125, 1–41. https://doi.org/10.1016/j.compedu.2018.06.008

Alharbi, S., & Drew, S. (2014). Using the technology acceptance model in understanding academics' behavioural intention to use learning management systems. International *Journal of Advanced Computer Science and Applications, 5*(1), 143–155. https://doi.org/10.14569/IJACSA.2014.050120

Bauwens, R., Muylaert, J., Clarysse, E., Audenaert, M., & Decramer, A. (2020). Teachers' acceptance and use of digital learning environments after hours: Implications for work-life balance and the role of integration preference. *Computers in Human Behavior, 112,* 106479. https://doi.org/10.1016/j.chb.2020.106479

Bennett, S., Dawson, P., Bearman, M., Molloy, E., & Boud, D. (2017). How technology shapes assessment design: Findings from a study of university teachers. *British Journal of Educational Technology, 48*, 672–682. *https://doi.org/10.1111/bjet.12439*

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PloS one*, *10*(12), e0143616. https://doi.org/10.1371/journal.pone.0143616

Born, S., Fink, A., Spoden, C., & Frey, A. (2019). Evaluating different equating setups in the continuous item pool calibration for computerized adaptive testing. *Frontiers in Psychology*, *10*, 1277. https://doi.org/10.3389/fpsyg.2019.01277

Brady, M., Devitt, A., & Kiersey, R. A. (2019). Academic staff perspective on technology for assessment (TfA) in higher education: A systematic literature review. *British Journal of Educational Technology, 50*, 3080–3098. https://doi.org/10.1111/bjet.12742

Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466. https://doi.org/ 10.1037/0033-290905.3.456

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. https://doi.org/10.1080/10705510701301834

Chua, Y. P., & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior, 29,* 1889–1895. https://doi.org/10.1016/j.chb.2013.03.008

Cigdem, H., & Topcu, A. (2015). Predictors of instructors' behavioral intention to use learning management system: A Turkish vocational college example. *Computers in Human Behavior, 52*, 22–28. https://doi.org/10.1016/j.chb.2015.05.049

Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quaterly, 19*(2), 189–211. https://doi.org/10.2307/249688

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user accpetance of information technology. *MIS Quarterly, 13(3),* 319–339. https://doi.org/10.2307/249008

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management Science. 35*, 982–1003. https://doi.org/10.1287/mnsc.35.8.982

Deeley, S. J. (2018). Using technology to facilitate effective assessment for learning and feedback in higher education. *Assessment & Evaluation in Higher Education, 43*, 439–448. https://doi.org/10.1080/02602938.2017.1356906

del Barrio-García, S., Arquero, J. L., & Romero-Frías, E. (2015). Personal learning environments acceptance model: The role of need for cognition, e-learning satisfaction and students' perceptions. *Educational Technology & Society, 18(3)*, 129–141.

Dolan, R. P., & Burling, K. S. (2017). Computer-based testing in higher education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (2nd Ed., pp. 370–384). Routledge. https://doi.org/10.4324/9781315709307.ch24

Fathema, N., Shannon, D., & Ross, M. (2015). Exapnding the technology acceptance model (TAM) to examine faculty use of learning management systems (LMSs) in higher education. *Merlot, 11*(2), 210–232. https://jolt.merlot.org/Vol11no2/Fathema_0615.pdf

Fink, A., Born, S., Frey, A., & Spoden, C. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, 60, 327–346. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_3-2018_327-346.pdf

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison-Wesley.

Fluck, A. E. (2019). An international review of eExam technologies and impact. *Computers & Education, 132*, 1–15. https://doi.org/10.1016/j.compedu.2018.12.008

Frey, A. (in press). Computerized adaptive testing and multistage testing. *International Encyclopedia of Education, 4th Edition*. Sage.

Frey, A., & Fink, A. (in press). Controlling for item position effects when adaptive testing is used in Large-Scale Assessments. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative computer-based international large-scale assessments – foundations, methodologies and quality assurance procedures*. Springer.

Garcia, A. V. M., & Gomez, M. C. S. (2014). Predictive model of the intention to adopt Blended Learning in a university setting. *Universitas Psychologica, 13*(2), 601–614. https://doi.org/10.11144/Javeriana.UPSY13-2.mpia

Gerick, J., Vennemann, M., Eickelmann, B., Bos, W., & Mews, S. (2018). *ICILS 2013. Dokumentation der Erhebungsinstrumente der International Computer and Information Literacy Study 2013* [ICILS 2013 documentation of the instruments of the International Computer and Information Literacy Study]. Waxmann.

Granić, A., & Marangunić, N. (2019). Technology acceptance model in educational context: A systematic review. *British Journal of Educational Technology, 50*, 2572–2593. https://doi.org/10.1111/bjet.12864

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155–167. https://doi.org/10.1007/s11336-008-9099-3

Gu, L., Ling, G., Liu, O. L., Yang, Z., Li, G., Kardanova, E., & Loyalka, P. (2020). Examining mode effects for an adapted Chinese critical thinking assessment. *Assessment & Evaluation in Higher Education, 46*, 870–893. https://doi.org/10.1080/02602938.2020.1836121

Harris, K. M., Phelan, L., McBain, B., Archer, J., Drew, A. J., & James, C. (2016). Attitudes toward learning oral communication skills online: The importance of intrinsic interest and student-instructor differences. *Educational Technology Research and Development, 64*(4), 591–609. https://doi.org/10.1007/s11423-016-9435

Hsu, L. (2016). Examining EFL teachers' technological pedagogical content knowledge and the adoption of mobile-assisted language learning: a partial least square approach. *Computer Assisted Language Learning, 29,* 1287-1297. https://doi.org/10.1080/09588221.2016.1278024

Igbaria, M., & Iivari, J. (1995). The effects of self-efficacy on computer usage. *Omega, 23*(6), 587–605. https://doi.org/10.1016/0305-0483(95)00035-6.

King, W. R., & He, J. (2006). A meta-analysis of the technology acceptance model. *Information & Management, 43*(6), 740–755. https://doi.org/10.1016/j.im.2006.05.003

Klösel, R. (2018). *Entwicklung eines Erhebungsinstruments zu Hinderungsgründen bei der Implementation eines neuen Konzepts für Hochschulklausuren* [Development of an instrument to measure obstacles during the implementation of a new concept for university exams] [Unpublished master's thesis]. Friedrich Schiller University Jena.

Kwon, O., Choi, K., & Kim, M. (2007). User acceptance of context-aware services: Self-efficacy, user innovativeness and perceived sensitivity on contextual pressure. *Behavior* & Information Technology, 26(6), 483–498. https://doi.org/10.1080/01449290600709111

Lee, Y., Hsieh, Y., & Chen, Y. (2013). An investigation of employees' use of e-learning systems: applying the technology acceptance model. *Behavior and Information Technology, 32*(2), 173–189. https://doi.org/10.1080/0144929X.2011.577190

Lee, D. Y., & Lehto, M. R. (2013). User acceptance of YouTube for procedural learning: an extension of the technology acceptance model. *Computers & Education, 61,* 193–208. https://doi.org/10.1016/j.compedu.2012.10.001

Lee, C.-L., & Huang, M.-K. (2014). The influence of computer literacy and computer-anxiety on computer self-efficacy: The moderating effect of gender. *Cyberpsychology, Behavior, and Social Networking, 17*, 172–181. https://doi.org/10.1089/cyber.2012.0029

Lin, S., Zimmer, J. C., & Lee, V. (2013). Podcasting acceptance on campus: The differing perspectives of teachers and students. *Computers & Education, 68*, 416–428. https://doi.org/10.1016/j.compedu.2013.06.003

Liu, H., Wang, L., & Koehler, M. J. (2019). Exploring the intention-behavior gap in the technology acceptance model: A mixed-methods study in the context of foreign-language teaching in China. *British Journal of Educational Technology, 50,* 2536–2556. https://doi.org/10.1111/bjet.12824

Mac Callum, K., Jeffrey, L., & Kinshuk (2014). Comparing the role of ICT literacy and anxiety in the adoption of mobile learning. *Computers in Human Behavior, 39,* 8–19. http://doi.org/10.1016/j.chb.2014.05.024

Maican, C. I., Cazan, A.-M., Lixandroiu, R. C., & Dovleac, L. (2019). A study on academic staff personality and technology acceptance: The case of communication and collaboration applications, *Computers & Education, 128,* 113–131, https://doi.org/10.1016/j.compedu.2018.09.010.

Maqableh, M., Masa'deh, R., & Mohammed, A. B. (2015). The acceptance and use of computer based assessment in higher education. *Journal of Software Engineering and Applications, 8*, 557–574. https://doi.org/10.4236/jsea.2015.810053

Marangunić, N., & Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information Society, 14*(1), 81–95. http://doi.org/10.1007/s10209-014-0348-1

McGill, T., Klobas, J., & Renzi, S. (2011). LMS use and Instructor Performance: The Role of Task-technology Fit. *International Journal on E-Learning, 10* (1), 43–62.

Mei, B., Brown, G. T. L., & Teo, T. (2018). Toward an Understanding of Preservice English as a Foreign Language Teachers' Acceptance of Computer-Assisted Language Learning 2.0 in the People's Republic of China. *Journal of Educational Computing Research, 56,* 74–104. https://doi.org/10.1177/0735633117700144

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*, 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Morris, L.W., Davis, M. A. & Hutchings, C. H. (1981). Cognitive and emotional components of anxiety: Literature review and a revised worry-emotionality scale. *Journal of Educational Psychology, 73*, 541–555. https://doi.org/10.1037/0022-0663.73.4.541

Motaghian, H., Hassanzadeh, A., & Moghadam, D. K. (2013). Factors affecting university instructors' adoption of web-based learning systems: case study of Iran. *Computers & Education, 61,* 158–167. https://doi.org/10.1016/j.compedu.2012.09.016

Mutambara, D., & Bayaga, A. (2021). Determinants of mobile learning acceptance for STEM education in rural areas. *Computers & Education, 160,* 104010. https://doi.org/10.1016/j.compedu.2020.104010

Muthén, L. K., and Muthén, B. O. (2018). *Mplus User's Guide, 8th Edn.* Los Angeles, CA: Muthén & Muthén.

Nardi, A., & Ranieri, M. (2018). Comparing paper-based and electronic multiple-choice examinations with personal devices: Impact in students' performance, self-efficacy and satisfaction. *British Journal of Educational Technology, 50,* 1495–1506. https://doi.org/10.1111/bjet.12644

Nikou, S. A., & Economides, A. A. (2016). The impact of paper-based, computer-based and mobile-based self-assessment on students' science motivation and achievement. *Computers in Human Behavior, 55,* 1241–1248. https://doi.org/10.1016/j.chb.2015.09.025

Nikou, S. A., & Economides, A. A. (2017). Mobile-based assessment: Investigating the factors that influence behavioral intention to use. *Computers & Education, 109,* 56–73. https://doi.org/10.1016/j.compedu.2017.02.005

Nikou, S. A., & Economides, A. A. (2018a). Mobile-based assessment: A literature review of publications in major referred journals from 2009 to 2018. *Computers & Education, 125,* 101–119. https://doi.org/10.1016/j.compedu.2018.06.006

Nikou, S. A., & Economides, A. A. (2018b). Factors that influence Behavioral Intention to Use Mobile-Based Assessment: a STEM teachers' perspective, *British Journal of Educational Technology*, *50,* 587–600. https://doi.org/10.1111/bjet.12609

Nistor, N. (2014). When technology acceptance models won't work: Non-significant intention-behavior effects. *Computers in Human Behavior, 34,* 299–300. https://doi.org/10.1016/j.chb.2014.02.052

Noble, S. M., Saville, J. D., & Foster, L. L. (2022). VR as a choice: what drives learners' technology acceptance?. *International Journal of Educational Technology in Higher Education, 19*, 6. https://doi.org/10.1186/s41239-021-00310-w

Paiva, J., Morais, C., Costa, L., & Pinheiro, A. (2017). The shift from "e-learning" to "learning": Invisible technology and the dropping of the "e". British Journal of *Educational Technology, 47*, 226–238. https://doi.org/10.1111/bjet.12242

Richter, T., Naumann, J., & Horz, H. (2010). Eine revidierte Fassung des Inventars zur Computerbildung (INCOBI-R) [A revised version of the Computer Literacy Inventory]. *Zeitschrift für Pädagogische Psychologie, 24,* 23–27. https://doi.org/10.1024/1010-0652/a000002

Rolim, C., & Isaias, P. (2019). Examining the use of e-assessment in higher education: Teachers and students' viewpoints. *British Journal of Educational Technology, 50,* 1785–1800. https://doi.org/10.1111/bjet.12669

Sánchez, R. A., & Hueros, A. D. (2010). Motivational factors that influence the acceptance of Moodle using TAM. *Computers in Human Behavior, 26*, 1632–1640. https://doi.org/10.1016/j.chb.2010.06.011

Sánchez-Prieto, J. C., Olmos-Migueláñez, S., & García-Peñalvo, F. J. (2016). Informal tools in formal contexts: Development of a model to assess the acceptance of mobile technologies among teachers. *Computers in Human Behavior, 55,* 519–528. https://doi.org/10.1016/j.chb.2015.07.002

Schepers, J., & Wetzels, M. (2007). A meta-analysis of the technology acceptance model: Investigating subjective norm and moderation effects. *Information & Management, 44(1)*, 90–103. https://doi.org/10.1016/j.im.2006.10.007

Scherer, R., Siddiq, F., & Teo, T. (2015). Becoming more specific. Measuring and modeling teachers' perceived usefulness of ICT in the context of teaching and learning. *Computers & Education, 88,* 202–214. https://doi.org/10.1016/j.compedu.2015.05.005

Scherer, R., Siddiq, F., & Tondeur, J. (2019). The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Computers & Education, 128,* 13–35. https://doi.org/10.1016/j.compedu.2018.09.009

Scherer, R., Siddiq, F. & Tondeur, J. (2020). All the same or different? Revisiting measures of teachers' technology acceptance. *Computers & Education, 143,* 103656. https://doi.org/10.1016/j.compedu.2019.103656

Scherer, R, & Teo, T. (2019). Unpacking teachers' intentions to integrate technology: A meta-analysis. *Educational Research Review, 27,* 90–109. https://doi.org/10.1016/j.edurev.2019.03.001

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *The encyclopedia of social measurement* (pp. 429–438). Elsevier/Academic.

Spoden, C., & Frey, A. (Eds.) (2021). *Psychometrisch fundierte E-Klausuren für die Hochschule [Psychometrically sound e-exams for higher education]*. Pabst Science Publishers.

St-Onge, C., Quellett, K., Lakhal, S., Dubé, T., & Marceau, M. (2021). COVID-19 as the tipping point for integrating e-assessment in higher education practices. *British Journal of Educational Technology, 53*, 349–366. https://doi.org/10.1111/bjet.13169

Szajna, B. (1996). Empirical evaluation of the revised technology acceptance model. *Management Science, 42,* 85–92. https://www.jstor.org/stable/2633017

Teo, T. (2009). Is there an attitude problem? Reconsidering the role of attitude in the TAM. *British Journal of Educational Technology, 40,* 1139–1141. https://doi.org/10.1111/j.1467-8535.2008.00913.x

Terzis, V,. & Economides, A. A. (2011). The acceptance and use of computer based assessment. *Computers & Education, 56,* 1032–1044. https://doi.org/10.1016/j.compedu.2010.11.017

Terzis, V., Moridis, C. N., & Economides, A. A. (2012). How student's personality traits affect Computer Based Assessment Acceptance: Integrating BFI with CBAAM. *Computers in Human Behavior, 28,* 1985–1996. https://doi.org/10.1016/j.chb.2012.05.019

van de Schoot, R., Lugtig, P., & Hox, J.J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*, 486–492. https://doi.org/10.1080/17405629.2012.686740

Venkatesh, V. (2000). Determinants of perceived ease of use: integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information System Research, 11*(4), 342–365. https://doi.org/10.1287/isre.11.4.342.11872

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: four longitudinal field studies. *Management Science, 46,* 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Venkatesh, V., & Morris, M. G. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. *MIS Quarterly, 24(1),* 115–139. https://doi.org/10.1287/mnsc.46.2.186.11926

Venkatesh, V., Morris, M., Davis, G., & Davis, F. (2003). User acceptance of information technology: towards a unified view. *MIS Quarterly, 27*(3), 479–501. https://doi.org/10.2307/30036540

Wang, W.-T., & Wang, C.-C. (2009). An empirical study of instructor adoption of web-based learning systems. *Computers & Education, 53*(3), 761–774. https://doi.org/10.1016/j.compedu.2009.02.021

Wu, J., & Du, H. (2012). Toward a better understanding of behavioral intention and system usage constructs. *European Journal of Information Systems, 21*(6), 680–698. https://doi.org/10.1057/ejis.2012.15

Yen, D. C., Wu, C.-S., Cheng, F.-F., & Huang Y.-W. (2010). Determinants of users' intention to adopt wireless technology: an empirical study by integrating TTF and TAM. *Computers in Human Behavior, 26,* 906–915. https://doi.org/10.1016/j.chb.2010.02.005

Zheng, M. & Bender, D. (2019). Evaluating outcomes of computer-based classroom testing: Student acceptance and impact on learning and exam performance. *Medical Teacher, 41*(1), 75-82. https://doi.org/10.1080/0142159X.2018.1441984

Appendix A

Table A.1

Adapted Wording for the German Versions of the Scales used from ICILS 2013 (Translated into English by the Authors)

Original Wording	Adapted Wording
Schülerinnen und Schüler [secondary school students]	Studentinnen und Studenten [students in higher education]
Unterricht [class]	Lehrveranstaltungen [courses]
Unterrichtsstunden [lessons]	Sitzungen [sessions]
Schuljahr [school year]	Semester [semester]

Note. ICILS 2013 = International Computer and Information Literacy Study (Gerick et al., 2018).

Table A.2

Item Wording for the Four Self-Developed Scales (Translated into English by the Authors)

Variable	Item	Wording
PU	01	E-Klausuren haben das Potential meine Arbeit zu erleichtern. [E-exams have the potential to make my work easier.]
	02	E-Klausuren haben das Potential meine Effektivität zu steigern. [E-exams have the potential to increase my effectiveness.]
	03	E-Klausuren sind ein guter Weg, um die Prüfungsbelastung zu reduzieren. [E-exams are a good way to reduce the working load of examinations.]
	04	E-Klausuren erhöhen die Qualität von Hochschulklausuren. [E-exams increase the quality of university exams.]
	05	E-Klausuren erleichtern den Prüfungsvorgang. [E-exams facilitate the examination process.]
	06	E-Klausuren ermöglichen eine fairere Bewertung im Vergleich zu papierbasierten Klausuren. [E-exams make fairer evaluations possible compared to paper-based exams.]
	07	E-Klausuren erscheinen mir nützlich. [E-exams seem to be useful.]
	08	Die Nutzung von E-Klausuren eröffnet im Vergleich zu papierbasierten Klausuren bessere Möglichkeiten zur Überprüfung der Lernziele von Lehrveranstaltungen. [Compared to paper-based exams, the use of e-exams opens up better possibilities to evaluate the learning objectives of courses.]
	09	Durch die Verwendung von E-Klausuren kann ich im Vergleich zu papierbasierten Klausuren Zeit bezüglich der Auswertung und Benotung von Klausuren einsparen. [By using e-exams I can save time in the evaluation and grading of exams compared to paper-based exams.]
SN	01	Ich bin der Meinung, dass die Mehrheit der in der Lehre Tätigen an meiner Hochschule E-Klausuren für eine gute Sache halten. [I think that the majority of the teaching staff at my university considers e-exams to be a good thing.]
	02	Meine Kollegen haben mir bereits dazu geraten, E-Klausuren zu nutzen. [My colleagues have already advised me to use e-exams.]
	03	Meine Hochschule unterstützt die Nutzung neuer Technologien in der Lehre (wie z. B. E-Klausuren-Systeme). [My university supports the use of new technologies in teaching (such as, e.g., e-exam systems).]
	04	Meine Hochschule ist darauf bedacht, stets „up-to-date“ zu sein. [My university strives to be constantly up-to-date.]
	05	Meine Studentinnen und Studenten befürworten den Einsatz von E-Klausuren. [My students support the use of e-exams.]
FC	01	An meiner Hochschule stehen Ansprechpartnerinnen und Ansprechpartner zur Verfügung, die mich bei der Entwicklung und Umsetzung von E-Klausuren unterstützen können. [At my university, there are contact persons who can support me in the development and implementation of e-exams.]

Variable	Item	Wording
	02	Mir sind Fort-/ und Weiterbildungsmöglichkeiten zum Thema E-Klausuren an meiner Hochschule bekannt. [I am aware of training opportunities on the subject of e-exams at my university.]
	03	Die technischen Ressourcen meiner Hochschule (z. B. Computerpools, Intranet) sind zur Durchführung von E-Klausuren geeignet. [The technical resources at my university (e.g., computer pools, intranet) are suitable for running e-exams.]
	04	Ich weiß, an wen ich mich innerhalb meiner Hochschule bei Fragen oder Problemen in Bezug auf E-Klausuren wenden kann. [I know the contact person at my university for questions or problems regarding e-exams.]
	05	Der Einsatz von E-Klausuren ist bereits in der Prüfungsordnung meiner Hochschule geregelt. [The use of e-exams is already regulated in the examination regulations of my university.]
	06	Meine Hochschule fördert den Einsatz von E-Klausuren durch die Bereitstellung entsprechender Ressourcen. [My university promotes the use of e-exams by providing appropriate resources.]
BI	01	Ich beabsichtige innerhalb der nächsten 2 Jahre E-Klausuren durchzuführen. [I intend to use e-exams within the next 2 years.]
	02	Ich werde E-Klausuren nur einsetzen, wenn ich dazu verpflichtet werde. [I will only use e-exams if I am obliged to do so.]*
	03	Ich werde E-Klausuren einsetzen, wenn ich entsprechende Unterstützung von Fachleuten erhalte. [I will use e-exams if I receive appropriate support from experts.]
	04	Ich werde E-Klausuren einsetzen, wenn die entsprechenden technischen Voraussetzungen gegeben sind. [I will use e-exams if the appropriate technical requirements are met.]
	05	Ich habe großes Interesse an der Nutzung von E-Klausuren. [I am very interested in using e-exams.]
	06	Ich werde E-Klausuren zukünftig papierbasierten Klausuren vorziehen. [In the future, I will prefer e-exams to paper-based exams.]
	07	Ich werde künftig auf E-Klausuren, wenn irgend möglich, verzichten. [In the future, I will avoid e-exams if at all possible.]*
	08	Ich plane künftig E-Klausuren zu verwenden. [I plan to use e-exams in the future.]

Note. * Inverted scoring. PU = Perceived usefulness, SN = Subjective norm; FC = Facilitating conditions; BI = Behavioral intention to use. Each item of these scales had a four-point Likert scale ranging from 1 *trifft gar nicht zu* [totally disagree] to 4 *trifft völlig zu* [totally agree].

Appendix B

Table B.1

Results for the Measurement Model

	Group 1 (e-exams)					Group 2 (adaptive e-exams)				
Factor Items	*M*	SD	λ	*ω*	*AVE*	*M*	SD	λ	*ω*	*AVE*
BI	2.41	0.82		.881	.714	2.22	0.77		.916	.680
BI01			.873					.953		
BI02			.790					.659		
BI03			.662					.807		
BI04			.814					.875		
BI05			.958					.940		
BI06			.853					.781		
BI07			.857					.710		
BI08			.920					.825		
PU	2.36	0.72		.874	.620	2.05	0.76		.940	.710
PU01			.917					.904		
PU02			.871					.898		
PU03			.760					.811		
PU04			.756					.834		
PU05			.727					.820		
PU06			.681					.832		
PU07			.943					.927		
PU08			.671					.775		
PU09			.706					.768		
PE	1.67	0.46		.858	.524	1.68	0.45		.809	.523
PE01			.765					.754		
PE02			.668					.696		
PE03			.640					.614		
PE04			.703					.722		
PE05			.721					.738		
PE06			.743					.716		
PE07			.748					.756		
PE08			.644					.659		
PE09			.790					.743		
PE10			.754					.755		
PE11			.798					.775		
PE12			.697					.768		
PE13			.768					.721		
PE14			.675					.678		
CSE[a]	2.65	0.32		.853	.571	2.68	0.32		.835	.567
CSE05			.833					.905		
CSE06			.748					.736		

	Group 1 (e-exams)					Group 2 (adaptive e-exams)				
Factor Items	*M*	SD	λ	*ω*	*AVE*	*M*	SD	λ	*ω*	*AVE*
CSE07			.730					.620		
CSE10			.758					.727		
CSE11			.675					.655		
CSE12			.930					.926		
CSE13			.631					.689		
CSE14			.698					.700		
CA	1.50	0.52		.871	.585	1.48	0.50		.832	.582
CA01			.799					.880		
CA02			.780					.684		
CA03			.694					.696		
CA04			.732					.749		
CA05			.794					.849		
CA06			.750					.790		
CA07			.871					.759		
CA08			.679					.669		
SN	2.11	0.48		.717	.500	2.19	0.53		.722	.479
SN01			.720					.634		
SN02			.671					.710		
SN03			.706					.748		
SN04			.619					.617		
SN05			.807					.741		
FC	2.00	0.72		.873	.683	2.12	0.82		.902	.668
FC01			.863					.897		
FC02			.863					.846		
FC03			.651					.656		
FC04			.873					.865		
FC05			.819					.750		
FC06			.868					.863		

Note. λ = Standardized factor loading; ω = Green and Yang's (2009) variation of reliability coefficient ω; *AVE* = Average variance extracted; BI = Behavioral intention to use; PU = Perceived usefulness; PE = Prior experience; CSE = Computer self-efficacy; CA = Computer anxiety; SN = Subjective norm; FC = Facilitating conditions.

[a] Six items were removed due to limited variance.

Appendix C

Table C.1

Summary of Hypotheses Testing

H	Path	Group 1(e-exams)				Group 2 (adaptive e-exams)			
		β	*SE*(β)	*p*	As expected	β	*SE*(β)	*p*	As expected
1	PU→BI	.782	0.020	≤.001	Yes	.879	0.008	≤.001	Yes
2	PE→PU	.241	0.055	≤.001	Yes	.259	0.058	≤.001	Yes
3	CSE→PE	.453	0.049	≤.001	Yes	.428	0.052	≤.001	Yes
4	CSE→PU	.080	0.065	.220	No	-.011	0.058	.843	No
5	CSE→BI	.041	0.041	.204	No	.087	0.029	.002	Yes
6	CA→CSE	-.502	0.048	≤.001	Yes	-.482	0.042	≤.001	Yes
7	CA→PE	-.034	0.042	.418	No	-.049	0.034	.150	No
8	SN→PU	.521	0.032	≤.001	Yes	.238	0.053	≤.001	Yes
9	SN→BI	.171	0.171	≤.001	Yes	.152	0.052	.003	Yes
10	FC→BI	.030	0.040	.460	No	.064	0.061	.329	No

Note. H = Hypothesis; β = Standardized path coefficients; *SE* = Standard error; BI = Behavioral intention to use; PU = Perceived usefulness; PE = Prior experience; CSE = Computer self-efficacy; CA = Computer anxiety; SN = Subjective norm; FC = Facilitating conditions.

Anhang D: Beitrag 4 – Kriteriumsorientiertes adaptives Testen mit der KAT-HS-App

Zitation: Fink, A., Spoden, C., Frey, A. & Naumann, P. (2021). Kriteriumsorientiertes adaptives Testen mit der KAT-HS-App. *Diagnostica, 67*(2)*,* 110–114. http://doi.org/10.1026/00121924/a000268

Softwareinformation

Kriteriumsorientiertes adaptives Testen mit der KAT-HS-App

Aron Fink[1], Christian Spoden[2], Andreas Frey[1,3] und Patrick Naumann[1]

[1]Institut für Psychologie, Goethe-Universität Frankfurt, Frankfurt am Main
[2]Deutsches Institut für Erwachsenenbildung, Leibniz-Zentrum für Lebenslanges Lernen e. V., Bonn
[3]Centre for Educational Measurement (CEMO), University of Oslo, Norwegen

Zusammenfassung: In dieser Softwareinformation werden die Möglichkeiten zur Konstruktion, Administration und Auswertung kriteriumsorientierter, computerisierter adaptiver und nicht-adaptiver Tests mit der R-basierten open-source KAT-HS-App erläutert. Die App ermöglicht unter anderem auch die Anwendung der kontinuierlichen Kalibrierungsstrategie von Fink, Born, Spoden und Frey (2018).

Schlüsselwörter: R-Software, computerbasiertes Testen, computerisiertes adaptives Testen, kontinuierliche Kalibrierung, Testentwicklung′

Criterion-Referenced Adaptive Tests Using the KAT-HS App

Abstract: This software demonstration presents the possibilities for the construction, administration, and evaluation of criterion-referenced, computerized adaptive and nonadaptive tests with the R-based open-source KAT-HS app. This app enables users to apply the continuous item calibration strategy of Fink, Born, Spoden, and Frey (2018).

Keywords: R software, computer-based testing, computerized adaptive testing, continuous item calibration, test development

Die Nutzung digitaler Technologien im Rahmen der psychologischen Diagnostik ermöglicht den Einsatz innovativer Itemformate und hat das Potential die Effizienz des Testprozesses erheblich zu steigern. Zudem bietet der Einsatz computerbasierter Verfahren die Möglichkeit, verschiedene leistungsfähige psychometrische Methoden in die Testpraxis zu integrieren. Besonders hervorzuheben sind hier Methoden auf Basis von Modellen der Item-Response-Theory (IRT; z. B. van der Linden, 2016). Die einzelnen Methoden sind gut untersucht und dokumentiert. Abseits der Grundlagenforschung und Anwendungen bei groß angelegten Vergleichsstudien wie PISA, IGLU oder TIMSS findet man jedoch nach wie vor kaum IRT-basierte Tests, obgleich sie das Potential haben in zahlreichen weiteren Anwendungsbereichen die Qualität von Testungen erheblich zu steigern. Ein Beispiel für einen solchen Anwendungsbereich mit hohem Testaufkommen sind Hochschulklausuren. Hier existiert eine deutliche Lücke zwischen psychometrischem Kenntnisstand und Testpraxis. Problematisch an derzeitigen Hochschulklausuren sind nach Spoden, Frey, Fink und Naumann (2020) im Wesentlichen vier Aspekte: 1. Lernziele werden nicht angemessen durch die genutzten Aufgaben operationalisiert. 2. Klausuren sind nicht als kriteriumsorientierte Verfahren konzipiert, so dass Ergebnisse nicht als Ausmaß des Erreichens von Lernzielen interpretiert werden können. 3. Testzeitpunkte werden nicht statistisch verlinkt, so dass die Unabhängigkeit der Ergebnisse von Kohortenleistungsfähigkeit und Klausurschwierigkeit nicht gewährleistet ist. 4. Die Messpräzision schwankt über den Merkmalsbereich mit typischerweise deutlich niedrigerer Messpräzision an den Rändern der Kompetenzverteilung. Spoden und Frey (im Druck) beschreiben in ihrem Konzept psychometrisch fundierter Hochschulklausuren, wie diesen Problemen durch die zielgerichtete Nutzung IRT-basierter Methoden begegnet werden kann. Ausgehend von diesem Konzept wurde die hier vorgestellte KAT-HS-App (KAT-HS = kriteriumsorientiertes adaptives Testen in der Hochschule) entwickelt. In ihrer Anwendbarkeit ist die App allerdings nicht auf Hochschulklausuren beschränkt. Vielmehr hat sie zum Ziel, IRT-basierte Methoden für neue Anwendungsbereiche zu erschließen. Hierfür werden verschiedene Elemente, für deren Anwendung normalerweise unterschiedliche Softwarepakete einzusetzen sind, in einem Programm gebündelt. Ihre Kernelemente sind: a) computerbasierte Testadministration; b) IRT-Skalierung; c) Methoden zur Überprüfung der psychometrischen Qualität des Tests; d) Online-Kalibrierung bei wiederholten Anwendungen eines Tests; e) Kontrolle von Itempositionseffekten (IPE);

Diagnostica (2021), 67 (2), 110–114
https://doi.org/10.1026/0012-1924/a000268

© 2021 The Author(s) Distributed as a Hogrefe OpenMind article
under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

f) computerisiertes adaptives Testen (z. B. Frey, 2020); und g) kriteriumsorientiertes Testen. Zudem setzt die App auf eine intuitive Benutzerführung mit sinnvollen Voreinstellungen, die es dem Anwender möglichst einfach macht IRT-basierte Tests auf professionelle Weise einzusetzen. Das Ziel ist somit die Zugänglichkeit IRT-basierter Methoden zu verbessern und gleichzeitig die methodisch angemessene Nutzung sicherzustellen. Die psychometrischen Grundlagen werden im Konzept von Spoden und Frey (im Druck) beschrieben, zu dem ein Workshop kostenfrei als Video abgerufen werden kann (https://kat-hs.uni-frankfurt.de/materialien/workshop/). Die KAT-HS-App ist eine in R programmierte Shiny-App (Chang, Cheng, Allaire, Xie & McPherson, 2019) und somit kostenfrei und open-source. Sie greift neben eigens programmierten Funktionen auf Routinen der R-Pakete mirt (Chalmers, 2012), mirtCAT (Chalmers, 2016) und equateIRT (Battauz, 2015) zurück. Die App ist nach Registrierung über die Website https://kat-hs.uni-frankfurt.de/materialien/software/ für Forschung und Lehre kostenfrei erhältlich. In der jetzigen Form ist die App ausschließlich für die Verwendung auf Windowssystemen geeignet. Im Folgenden werden die grundlegenden Funktionalitäten der App vorgestellt. Eine ausführlichere Dokumentation der KAT-HS-App ist im zugehörigen Benutzerhandbuch zu finden.

Testkonstruktion

Die Nutzung der KAT-HS-App setzt einen Itempool voraus, der ein eindimensionales Merkmal inhaltsvalide abbildet. Die Items sollen zudem konform mit den Annahmen des verwendeten IRT-Modells sein. Inwieweit dies zutrifft, wird bei der Nutzung der App berechnet und Vorschläge zum Umgang mit nicht hinreichend fittenden Items gegeben. Die Iteminformationen (z. B. Item-ID, Stimulus, Antwortoptionen, Inhaltsbereich, etc.) müssen in einer Itemdatenbank im XLSX-Format zusammengestellt werden. Durch die Möglichkeit, Itemstämme und etwaige Antwortoptionen in HTML-Code zu spezifizieren, existieren zahleiche Möglichkeiten der Itemgestaltung. So ist beispielsweise auch das Einbinden von Multimediadateien möglich. Es können automatisch auswertbare und von Hand zu kodierende Antwortformate genutzt werden. Die App kann nur mit dichotom bewerteten Items arbeiten. Existierende Itemparameterschätzungen können zu Beginn in der Itemdatenbank eingefügt oder bei wiederholten Testungen im laufenden Betrieb durch Nutzung der kontinuierlichen Kalibrierungsstrategie (KKS; Fink, Born, Spoden & Frey, 2018; Frey & Fink, in press) ergänzt werden.

Nach Import der Itemdatenbank können verschieden komplexe Testarten mit der App erstellt werden. Vom einfachen linearen Test mit nur einer Testversion bis zum volladaptiven Test mit KKS inklusive Kontrolle von IPE und kriteriumsorientierter Testwertinterpretation ergeben sich zahlreiche Möglichkeiten der Testzusammenstellung. Hierfür können in der App noch weitere Einstellungen getroffen werden (z. B. Testlänge, Bearbeitungszeit, Anzahl Itemcluster, Itemauswahlkriterium, Personenparameterschätzer, Content-Balancing, etc.). Für den Fall von wiederkehrenden Testungen ist die KKS in die App implementiert. Mit dieser können Items im laufenden Testbetrieb ergänzt und kalibriert werden. Die KKS weist folgende Kernelemente auf: (a) Nutzung von Itemantworten mehrerer Testanwendungen zur Kalibrierung mit einem IRT-Modell, (b) Beibehaltung der Berichtsmetrik über Testungen, (c) ansteigende Adaptivität und Präzision der Fähigkeitsschätzungen über Testungen sowie (d) Kontrolle von Itemparameter Drift (IPD) und (e) IPE. In der KKS sind drei Arten von Itemclustern zu unterscheiden. Das *adaptive Cluster*, welches Items enthält, die im Testverlauf adaptiv gewählt werden und der Erhöhung der Messpräzision dienen; das *Kalibrierungscluster*, welches neue Items ohne Schätzung enthält, die der Vergrößerung des Itempools dienen; schließlich das *Linking-Cluster*, welches Items enthält, die zum Linking zweier aufeinander folgender Testzyklen genutzt werden. Die Auswahl der Linkitems kann anhand bestimmter Kriterien (z. B. Schwierigkeitsverteilung, Inhaltsbereich) automatisiert durch die App erfolgen. Je nach Charakteristika des Itempools, gewünschtem Grad an Adaptivität, Anzahl neuer Items, et cetera können in den Testeinstellungen mehr oder weniger der entsprechenden Cluster spezifiziert und so verschiedene Testarten erstellt werden. In Tabelle 1 sind die verschiedenen Testarten dargestellt, die durch die Kombination der drei Clusterarten konstruiert werden können.

Sind mehr als ein Itemcluster angegeben, erstellt die KAT-HS-App verschiedene Testversionen. Als Testdesign wird die Struktur eines balancierten lateinischen Quadrats (Williams, 1949) genutzt, um die Balancierung der Clusterpositionen und der Clusterreihenfolgen über Personen zu erreichen. Dies erhöht zum einen die Testsicherheit und ermöglicht zum anderen die Ausbalancierung von IPE (Frey, Bernhardt & Born, 2017) und Carry-Over-Effekten erster Ordnung auf der Ebene von Item-Clustern.

Zusätzlich muss eine Tabelle mit den Personeninformationen (z. B. Personen-ID, Login-Daten, Testversion) in die App importiert werden.

© 2021 The Author(s) Distributed as a Hogrefe OpenMind article
under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

Tabelle 1. Konfiguration unterschiedlicher Testarten mit der KAT-HS-App

Testart	Anzahl adaptiver Cluster	Anzahl Kalibrierungs-cluster	Anzahl Linking-cluster
Einzelne/erste Anwendung nicht-adaptiver Test	= 0	≥ 1	= 0
Wiederkehrende Anwendung nicht-adaptiver Test ohne Vergrößerung des Itempools	= 0	= 0	≥ 1
Verlinkung + Vergrößerung des Itempools	= 0	≥ 1	≥ 1
KKS ohne Vergrößerung des Itempools	≥ 1	= 0	≥ 1
KKS	≥ 1	≥ 1	≥ 1
Volladaptiver Test (Abbruchkriterium = Testlänge)	= 1	= 0	= 0

Anmerkung: KKS = Kontinuierliche Kalibrierungsstrategie.

Testadministration

Der Test kann im Standardbrowser des Testcomputers oder, für Szenarien die einen höheren Grad an Sicherheit benötigen, im Safe Exam Browser (ETH Zürich, 2019) gestartet werden. Letzterer versetzt den Computer in einen sogenannten Kioskmodus, welcher den Zugriff auf Hilfsmittel (z. B. externe Programme, Webseiten) einschränkt oder unterbindet. Der Safe Exam Browser muss vorher auf dem Testcomputer installiert und konfiguriert sein. Nach Beendigung des Tests speichert die Testanwendung die individuellen Ergebnisse als RData-Files. In diesen sind alle relevanten Informationen (bearbeitete Items, Rohantworten, kodierte Antworten, Bearbeitungszeiten etc.) gespeichert und können bei Bedarf in R eingesehen und weiterbearbeitet werden.

Auswertung

Schritt 1: Daten zusammenführen

Für die folgenden Auswertungsschritte benötigt die KAT-HS-App eine Antwortmatrix mit den dichotom kodierten Antworten der Testeilnehmer. Die App bietet die Möglichkeit zur automatisierten Erstellung einer solchen Matrix. Diese Antwortmatrix kann anschließend für die IRT-Skalierung genutzt werden. Zudem können die Antwortmatrizen aufeinander folgender Testanwendungen zusammengeführt sowie eine Tabelle mit den Rohantworten der Testpersonen exportiert werden.

Schritt 2: IRT-Skalierung

Die Skalierung kann in der KAT-HS-App mit oder ohne Verankerung auf frühere Testzeitpunkten (freie und fixierte Skalierung) erfolgen. Die fixierte Skalierung ist für Testdesigns mit Linkitems geeignet. Hierbei werden die Parameterschätzungen der Linkitems auf die Werte des vorangegangenen Testzyklus fixiert und alle anderen Itemparameter frei geschätzt. So werden individuelle Testergebnisse auf derselben Metrik verortet und sind über die verschiedenen Testzeitpunkte hinweg direkt vergleichbar. Als Messmodell können die geläufigen logistischen Testmodelle mit einem (1PL) oder zwei Parametern (2PL) genutzt werden (van der Linden, 2016). Die Schätzung der Itemparameter erfolgt mit dem bei den meisten aktuellen IRT-Programmen genutzten Marginal-Maximum-Likelihood-Verfahren (Bock & Aitken, 1981). Es ist über einen breiten Anwendungsbereich einsetzbar, tätigt vergleichsweise wenige Annahmen, erlaubt die konsistente Itemparameterschätzung von ein- und mehrparametrischen IRT-Modellen und ist auch für unvollständige Kalibrierungsdesigns (wie z. B. unter Verwendung der KKS) geeignet.

Schritt 3: Überprüfung der psychometrischen Qualität der Skalierung

Die Skalierungsergebnisse können hinsichtlich verschiedener Kennwerte auf ihre psychometrische Güte überprüft werden. Die App bietet die Möglichkeit Modellfitstatistiken, Itemstatistiken der klassischen Testtheorie sowie IRT-basierte Itemfitstatistiken zu berechnen. Für wiederkehrende Testanwendungen mit Linkitems können diese zudem auf IPD überprüft werden. Dafür werden die Itemparameterschätzungen aus den Skalierungen zweier aufeinanderfolgender Testanwendungen durch Equating-Methoden (z. B. Born, Fink, Spoden & Frey, 2019) auf eine gemeinsame Skala gebracht und mittels Wald-Test auf IPD getestet. Der Test auf IPD erfolgt iterativ. Zudem gibt es die Möglichkeit RDS-Files aus der App zu exportieren und für weitere Analysen in R zu nutzen.

Schritt 4: Personenparameterschätzung

Basierend auf den Itemparameterschätzungen können nun die Personenparameter anhand verschiedener Schätzverfahren (z. B. Glas, 2016) bestimmt werden. Für die grafische Darstellung der Item- und Personenparameter gibt es

© 2021 The Author(s) Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

die Möglichkeit eine Wright Map zu erstellen. Darüber hinaus wird die Reliabilität berechnet.

Schritt 5: Kategorisierung von Testergebnissen
In einigen Anwendungsbereichen ist es nützlich, Testergebnisse nicht nur als numerische Werte, sondern auch in Form inhaltlich definierter Kategorien zurückzumelden. Dies kann kriteriumsorientierte Testwertinterpretationen erleichtern. Bei vorliegenden Grenzwerten auf dem latenten Merkmalskontinuum können die Kategorisierungen mit der App automatisch durchgeführt werden. Wichtig im Hinblick auf die Validität der abgeleiteten Kategorieinterpretationen ist eine sorgfältige Bestimmung der Grenzwerte, die bei Kompetenztests üblicherweise auf Standard-Setting-Prozeduren basiert.

Dokumentation des Tests

Mit Hilfe des R-Paketes knitr (Xie, 2015) bietet die App die Möglichkeit, automatisiert eine Dokumentation des Tests zu erstellen (vgl. Spoden & Buchwald, 2018). Hierbei werden alle relevanten Informationen zum Test zusammengestellt und als DOCX-Dokument exportiert.

Zusammenfassung und Ausblick

Mit der KAT-HS-App liegt eine kostenfreie Software zur Konstruktion, Administration und Auswertung psychometrisch fundierter, computerbasierter Tests vor. Sie verfügt über eine grafische Nutzeroberfläche und erleichtert so auch Nutzergruppen ohne Programmierkenntnisse in R die Bedienung. Darüber hinaus werden mit der Installation ein Benutzerhandbuch, Vorlagen für Item- und Personentabellen sowie Beispieldatensätze bereitgestellt. Da die App sowohl die computerbasierte Testadministration, als auch die Analysen in einem Programm ermöglicht, füllt sie eine relevante Lücke innerhalb verfügbarer Software. Weiterhin ist sie die erste Software, mit der die für viele Anwendungsbereiche attraktive KKS direkt genutzt werden kann. Durch die Konzentration auf zentrale IRT-basierte Verfahren, der für viele Anwendungsbereiche geeigneten Voreinstellung an Methoden (die bei Bedarf angepasst werden kann) und einer einfachen Benutzerführung eröffnet sie Anwendern die Nutzung von IRT-basierten Tests, für die die bislang verfügbaren Softwarepakete eine zu hohe Hürde darstellten. Sollten versierte Nutzer bisher nicht implementierte Methoden anwenden wollen, können R-Objekte exportiert und für weiterführende Analysen in R genutzt werden. Künftig soll der Funktionsumfang der App erweitert werden (z. B. ordinale IRT-Modelle; Person-Fit). Die Weiterentwicklung soll durch das Bilden einer aktiven Nutzercommunity über das Onlineportal https://kat-hs.uni-frankfurt.de gefördert werden.

Literatur

Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software, 68(7),* 1 – 22. https://doi.org/10.18637/jss.v068.i07

Bock, R. D. & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46,* 443 – 459. https://doi.org/10.1007/BF02293801

Born, S., Fink, A., Spoden, C. & Frey, A. (2019). Evaluating different equating setups in the continuous item pool calibration for computerized adaptive testing. *Frontiers in Psychology*, *10*, 1277. https://doi.org/10.3389/fpsyg.2019.01277

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48(6),* 1 – 29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71(5),* 1 – 39. https://doi.org/10.18637/jss.v071.i05

Chang, W., Cheng, J. Allaire, J. J., Xie, Y. & McPherson, J. (2019). *Shiny: Web Application Framework for R* (R package version 1.3.2) [Computer Software].

ETH Zürich, Lehrentwicklung und -technologie (2019). *Safe Exam Browser* (Version 2.2.3) [Computer Software]. Zürich: ETH Zürich, Lehrentwicklung und -technologie (LET).

Fink, A., Born, S., Frey, A. & Spoden, C. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling, 60,* 327 – 346.

Frey, A. (2020). Computerisiertes adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3., aktual. und überarb. Auflage, S. 501 – 525). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-61532-4_20

Frey, A., Bernhardt, R. & Born, S. (2017). Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests. *Diagnostica, 63,* 167 – 178. https://doi.org/10.1026/0012-1924/a000173

Frey, A. & Fink, A. (in press). Controlling for item position effects when adaptive testing is used in Large-scale assessments. In L. Khorramdel, M. von Davier & K. Yamamoto (Eds.), *Innovative computer-based international Large-Scale Assessments – foundations, methodologies, and quality assurance procedures.* New York, NY: Springer.

Glas, C. A. W. (2016). Maximum-likelihood estimation. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume two: statistical tools* (pp. 197 – 216). London: Chapman and Hall.

Spoden, C. & Buchwald, F. (2018). Diagnostische Tests mit R und knitr: Erstellung, Auswertung und Vorbereitung der Rückmeldung. *Diagnostica, 64*, 49 – 57. https://doi.org/10.1026/0012-1924/a000189

Spoden, C. & Frey, A. (Hrsg.). (im Druck). *Psychometrisch fundierte E-Klausuren für die Hochschule.* Lengerich: Pabst Science Publishers.

© 2021 The Author(s) Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)

Spoden, C., Frey, A., Fink, A. & Naumann, P. (2020). Kompetenzorientierte elektronische Hochschulklausuren im Studium des Lehramts. In K. Kaspar, M. Becker-Mrotzeck, J. Hofhues, J. König & D. Schmeinck (Hrsg.), *Bildung, Schule und Digitalisierung* (S. 184 – 189). Münster: Waxmann.

Linden, W. J. van der (2016). *Handbook of item response theory, volume one: models.* London: Chapman and Hall. https://doi.org/10.1201/9781315374512

Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry, 2,* 149 – 168.

Xie, Y. (2015). *Dynamic documents with R and knitr* (The R series, 2nd ed.). Boca Raton, FL: Chapman & Hall / CRC.

Interessenskonflikt

Der korrespondierende Autor erklärt im Namen aller Autoren, dass kein Interessenkonflikt vorliegt.

Förderung

Die in dem Artikel berichtete Forschung wurde durch das Bundesministeriums für Bildung und Forschung (Ref: 16DHL1005) gefördert.

ORCID

Aron Fink
https://orcid.org/0000-0003-0624-1131

Aron Fink, M.Sc.
Goethe-Universität Frankfurt
Theodor-W.-Adorno-Platz 6
60323 Frankfurt am Main
a.fink@psych.uni-frankfurt.de

© 2021 The Author(s) Distributed as a Hogrefe OpenMind article under the license CC BY 4.0 (https://creativecommons.org/licenses/by/4.0)