Messung komplexer Konstrukte mit computerisierten Verfahren am Beispiel von ICT-Skills

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 Sonja Franziska Christina Wenzel

Geboren am 15.02.1984 in Finsterwalde, Niederlausitz

Frankfurt am Main

Mai 2020

Gutachter:	1.	Prof. Dr. Andreas Frey
	2.	Prof. Dr. Holger Horz

Datum der Abgabe: 13.05.2020

Datum der Disputation: 25.11.2020

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die mich bei der Erstellung dieser Arbeit unterstützt haben. Danke an

Andreas für die Möglichkeit diese Arbeit zu schreiben, für deine Unterstützung bei der Verfolgung meiner Forschungsideen, die Besprechungen wann immer sie für mich nötigen waren, die kritischen Anmerkungen, die wertvollen Verbesserungsvorschläge und das stetige Weiterdenken. All das hat mich fachlich und persönlich weitergebracht.

Holger für das Vertrauen, das du in mich gesetzt hast und die Geduld, die du bewiesen hast. Du hast mich stets motiviert und persönlich unterstützt. Manchmal hast du mir auch mehr zugetraut als ich mir selbst, danke dafür.

Das CavE-ICT-Projektteam – besonders Lena – dafür, dass wir über die Projektzeit hinaus im Kontakt sind, das möchte ich nicht missen. Meine Hilfskraft Monja, ohne dich und dein Engagement hätte es viele ICT-Items so nicht gegeben.

Meine Kollegen aus der Zeit, die ich an der FSU Jena verbracht habe, dafür wie viel ich in dieser Zeit mit euch und von euch lernen durfte. Vor allem an Sebastian für den oft langen und detaillierten Austausch zu allem, was mit CAT und MAT zu tun hat.

Meine PsyLLiE-Kollegen für die Anteilnahme und meiner Hilfskraft Maren, die mir viele „Kleinigkeiten" abgenommen hat.

Janine für das Lesen meiner Studien und die stundenlangen R-Sessions, die die allerschönsten Grafiken hervorgebracht haben, und an Daniel für die geduldige Überprüfung der Formalisierungen in Studie III.

Ganz besonders möchte ich meinen Mädels aus dem Writing Club danken: Claudi, Julia, Sonja und Yasemin. Ihr seid eine riesen Unterstützung und unsere akademischen Erfahrungen zu teilen ist für mich unglaublich wertvoll. Darüber hinaus ist es einfach schön, wenn man jeden Tag auf der Arbeit seine Freunde treffen kann. Besonders herausheben muss ich an dieser Stelle Claudi – danke für die vielen fordernden, kritischen, aufbauenden und lustigen Kommentare an meinen Texten. Ich wünschte du wärst Energie Cottbus Fan, denn ich glaube nicht, dass man mit dir als Unterstützerin verlieren kann.

Ich danke meiner Familie – meinen Eltern, die mir nie zu- oder von etwas abgeraten haben, die mir aber ihre Meinung ehrlich sagen, immer an meiner Sicht interessiert sind, und mich bestärken Entscheidungen zu treffen, zu denen ich stehen kann. Meinem Großvater Rudolf, der mir vorgelebt hat, wie toll es ist, wenn man seinen Job lebt. Meiner Urgroßmutter Marianne, weil sie mir den Wert von Emanzipation und Frauenpower verdeutlicht hat, bevor ich überhaupt verstanden habe, dass dies wichtige Werte sind.

Meine Freunde, die wie selbstgewählte Familie für mich sind: Anne, Basti und Jule. Ihr habt diesen Prozess von Anfang bis Ende und mich durch Höhen und Tiefen begleitet. Ob wir in der gleichen Straße oder auf unterschiedlichen Kontinenten zu Hausen sind, ihr seid da – Danke!

Audere est Facere

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Um den aktuellen Bildungsstand einer Gesellschaft abbilden zu können müssen Resultate von Bildungsprozessen, wie erworbenes Wissen oder ausgebildete Fähigkeiten, modelliert und gemessen werden (Leutner, Klieme, Fleischer & Kuper, 2013). Im Rahmen sogenannter Large-Scale-Assessments (LSAs) werden Kompetenzen in bestimmten Bereichen definiert und erfasst, die generell für die gesellschaftliche Teilhabe benötigen werden (bspw. Fraillon, Schulz & Ainley, 2013). Durch die fortschreitende Digitalisierung aller Lebens- und Arbeitsbereiche ist der kompetente Umgang mit Informations- und Kommunikationstechnologien (ICT) eine wichtige Voraussetzung für die erfolgreiche Teilhabe an unserer modernen Wissensgesellschaft. Die detaillierte Beschreibung solcher, auch als ICT-Skills bezeichneter Kompetenzen, und die Entwicklung von theoriebasierten Instrumenten zu deren Erfassung ist von großer Bedeutung, um mögliche sozial bedingte Disparitäten aufzudecken.

Im Rahmen der vorliegenden Arbeit werden Annahmen, Ergebnisse und Daten aus dem Projekt CavE-ICT, in dem verhaltensnahe simulationsbasierte Items zur Erfassung von ICT-Skills entwickelt wurden, aufgegriffen und weitergenutzt mit dem Ziel eine besonders effiziente und ökonomisch Messung von ICT-Skills im LSA-Kontext und darüber hinaus zu ermöglichen. Ein vielversprechender Ansatz durch den Testzeiten verkürzt und/oder die Messpräzision erhöht werden kann ist das computerisierte adaptive Testen (CAT; bspw. Frey, 2012). Beim adaptiven Testen orientiert sich die Auswahl der Items am Antwortverhalten der untersuchten Person, so dass durch die Berücksichtigung der individuellen Fähigkeit einer Person Items mit möglichst viel diagnostischer Information administriert werden können. Damit auch bei der Vorgabe unterschiedlicher Items in unterschiedlicher Reihenfolge Testleistungen von Personen miteinander verglichen werden können, stellen Modelle der Item-Response-Theorie (IRT; bspw. Hambleton & Swaminathan, 2010) die Basis der Anwendung von CAT dar.

Im Rahmen dieser Arbeit wurde untersucht, wie ICT-Skills auf Basis der Item-Response-Theorie und unter Einsatz computerisierter Messinstrumente erfasst werden können. Dabei setzten die empirischen Studien dieser Arbeit unterschiedliche Testformen um und an unterschiedlichen Punkten im Prozess der Testentwicklung an. Studie I setzt noch vor der Entwicklung von Items zur Messung von ICT-Skills an und zielt darauf ab Hinweise zum Umfang des zu erstellenden ICT-Itempools und zur Testlänge eines adaptiven Messinstruments bereitzustellen. Studie II baut direkt auf Studie I auf und nutzt die im Rahmen des Projekts CavE-ICT entwickelten und kalibrierten Items beziehungsweise ihre ermittelten Itemeigenschaften zur weiteren Erprobung verschiedener CAT-Algorithmen. Es werden Möglichkeiten aufgezeigt, wie multidimensionales adaptives Testen zur Messung von ICT-Skills gewinnbringend eingesetzt werden kann, und zudem eine differenzierte Messung auf Ebene der verschiedenen kognitiven Prozesse von ICT-Skills erlaubt. Dabei werden explizit Möglichkeiten exploriert Items die unterschiedliche kognitive Prozesse von ICT-Skills abbilden sequentiell geordnet und trotzdem adaptiv vorzulegen. Die durch Studie II erarbeiteten Erkenntnisse können insbesondere für die Erfassung von multidimensionalen Konstrukten oder facettierten Merkmalen in LSAs genutzt werden. Durch den Vergleich der Ergebnisse von Studie I und II ergeben sich zudem Implikationen für ein angemessenes Design von Simulationsstudien die insbesondere noch vor der eigentlichen Test- beziehungsweise Itementwicklung ansetzen. In Studie III werden lineare Kurztests zur Messung von ICT-Skills zusammengestellt. Durch die gezielte Auswahl geeigneter ICT-Items soll bei möglichst geringer Testzeit zugleich eine hohe Messgenauigkeit und Zuverlässigkeit realisiert werden. Die in Studie III manuell und automatisiert computerbasiert zusammengestellten Tests werden hinsichtlich des Einsatzes sowohl auf Populationsebene, im Sinne einschlägiger LSAs, als auch darüber hinaus für gruppen- und individualdiagnostische Zwecke evaluiert und Empfehlungen für den Kurztesteinsatz abgeleitet.

1 Einleitung

Innerhalb einer Gesellschaft tragen Bildungsprozesse nicht nur dazu bei, dass Individuen neue Erfahrungen, neues Wissen und neue Fähigkeiten erwerben (Alheit & Dausien, 2002), sondern der Erwerb und die Anwendung dieser Wissensinhalte bestimmen auch maßgeblich den gesellschaftlichen Wohlstand, sozialen Zusammenhalt sowie Entwicklungschancen (Klieme & Leutner, 2006). Doch erst die Modellierung und Messung der Resultate von Bildungsprozessen, wie beispielsweise erworbenes Wissen oder ausgebildete Fähigkeiten und Fertigkeiten, Erlauben einen Rückschluss auf den aktuellen Bildungsstand innerhalb einer Gesellschaft und spielen eine entscheidende Rolle bei der Optimierung von Bildungsprozessen sowie der Qualitätssicherung und -entwicklung des Bildungswesens (Leutner et al., 2013). Als wesentliche Instrumente des Bildungsmonitorings und wichtiger Bereich der empirischen Bildungsforschung sind Large-Scale-Assessments (LSAs) zu nennen. Im Rahmen solcher groß angelegten Vergleichsstudien werden typischerweise Kompetenzen in bestimmten Bereichen definiert, die generell für die gesellschaftliche Teilhabe benötigen werden (bspw. Fraillon et al., 2013). Kompetenzen die in LSAs abgebildet werden, gelten meist im Sinne eines lebenslangen Lernens als bedeutsam für schulische und berufliche Entwicklung; sie fördern darüber hinaus aber auch anschlussfähiges Lernen (Rost, Prenzel, Carstensen, Senkbeil & Groß, 2004). Zudem wird im Zuge von LSAs häufig der Frage nachgegangen, ob im Bildungssystem hinsichtlich der Ausprägung bestimmter Kompetenzen sozial bedingte Disparitäten zu beobachten sind (Gniewosz & Gräsel, 2015). Dabei stellt sich zunehmen auch die Frage wie Bildungsprozesse unter den Bedingungen des digitalen Wandels erfolgreich gestaltet werden können. Digitale Technologien sind mit dem alltäglichen Erleben und Handeln fest verbunden und nicht mehr daraus wegzudenken. So können beispielsweise Smartphones nicht nur dazu genutzt werden, um zu telefonieren oder Nachrichten zu verschicken, sondern auch um zu fotografieren, Einkäufe und

Zahlungen zu tätigen, Recherchen im Netz zu betreiben oder einfach nur den digitalen Kalender zu verwalten. Als Voraussetzung für eine erfolgreiche Teilhabe an modernen Wissensgesellschaften ist der kompetente Umgang mit Informations- und Kommunikationstechnologien (ICT) unverzichtbar. Die Bedeutung solcher Kompetenzen, Fertigkeiten und Fähigkeiten welche im Folgenden unter ICT-Skills subsummiert werden, liegt vor allem darin, dass diese in allen Lebensbereichen benötigt werden und Menschen, welche nicht auf sie zurückgreifen können, Benachteiligungen erfahren können. Die vorliegende Arbeit hat zum Ziel, Fertigkeiten und Fähigkeiten von Schülerinnen und Schülern in Bezug auf die Bearbeitung von ICT-Aufgaben, die nicht nur rein technisches Wissen sondern auch kognitive Fähigkeiten erfordern, ökonomisch und in großem Maßstab, wie im Rahmen von LSAs erfassbar zu machen. Versteht man ICT-Skills als eine verhaltensbasierte Kompetenz, die sich durch die direkte Ausübung an einem Computer und die Interaktion mit Computerapplikationen auszeichnet, können ICT-Skills auch am besten durch computerbasierten Verfahren gemessen werden (Frey & Hartig, 2013). Dieser Logik folgend wird der Entwicklung eines Instruments zur Erfassung von ICT-Skills die zentrale Annahme zugrunde gelegt, dass nur mit direkt am Computer zu bearbeitenden und auf Simulationen basierenden Items, Verhaltensunterschiede gemessen werden können, die die interessierende Kompetenz direkt repräsentieren. So kann der tatsächliche Umgang mit ICT in der Testsituation geprüft werden und eine, zentral für die valide Interpretation von Testwerten, sehr gute Konstruktrepräsentation ermöglicht werden (Sireci & Zenisky, 2006; Goldhammer, Kröhne, Keßel, Senkbeil & Ihme, 2014). Eine theoretische Konzeptionen zur detaillierten Beschreibung von ICT-Skills sowie die Entwicklung von theoriebasierten Instrumenten beziehungsweise Testitems zu deren Erfassung sind daher von großer Bedeutung, da erst durch die konkrete Definition komplexer Kompetenzkonstrukte und die Formulierung

von theoretischen Kompetenzmodellen die Grundlage für deren empirische Untersuchung geschaffen wird.

Gerade im Rahmen von LSAs führen die breite Definition der untersuchten Kompetenzbereiche und deren Komplexität dazu, dass eine große Anzahl von Testitems für jeden Kompetenzbereich zur Verfügung steht (Walter & Rost, 2011). Insofern besteht in LSAs grundsätzlich das Dilemma zwischen der Abbildung einer möglichst breiten inhaltlichen Erfassung der Kompetenzbereiche und dem Wunsch nach Effizienz in der Testgestaltung, sodass vielen Probanden trotz begrenzter Testzeit möglichst viele Items vorgegeben werden können (Baumert, Stanat & Demmrich, 2001). Um den Testaufwand zu begrenzen sowie die Beanspruchung der Testteilnehmer zumutbar zu gestalten und Ermüdung und Leistungsrückgänge durch Demotivation zu vermeiden, ist nach Wegen zu suchen, die Testungen effizient zu gestalten (Frey & Seitz, 2010). Einen vielversprechenden Ansatz dazu bietet das computerisierte adaptive Testen (CAT; bspw. Frey, 2012). Mit einem adaptiven Test lassen sich Testsitzungen deutlich verkürzen und/oder die Messpräzision der gemessenen Kompetenzen deutlich steigern, da sich beim adaptiven Testen die Auswahl der Items am Antwortverhalten des untersuchten Individuums orientiert und durch die Berücksichtigung der individuellen Fähigkeit einer Person Items mit möglichst viel diagnostischer Information vorgelegt werden können. Damit auch bei der Vorgabe unterschiedlicher Items in unterschiedlicher Reihenfolge – wie es bei dem adaptiven Testen der Fall ist – Testleistungen von Personen problemlos miteinander verglichen werden können, stellen Modelle der Item-Response-Theorie (IRT; bspw. Hambleton & Swaminathan, 2010) die Basis der Anwendung von CAT dar (Frey, 2012). Wird eine Menge von Items mithilfe eines IRT-Modells kalibriert, können Personenfähigkeiten und spezifische Itemeigenschaften (Schwierigkeit, Diskrimination, Ratewahrscheinlichkeit) auf einer gemeinsamen Skala verortet werden und immer auf das zugrunde liegende Merkmal geschlossen werden – unabhängig davon, welche Items einer

Person vorgelegt wurden. Mitunter bieten unidimensionale IRT-Modelle, die nur eine Merkmalsausprägung der Person adressieren, keine ausreichende Grundlage zur Modellierung von Testergebnissen zu komplexen Konstrukten. Um die theoretischen Kompetenzmodelle in LSAs messen zu können, ist es meist notwendig, mehr als eine Merkmalsdimension in das Messmodell zu integrieren. Mit sogenannten multidimensionalen Modellen der Item-Response-Theorie (MIRT-Modellen) lassen sich prinzipiell die komplexen Strukturen von Kompetenzmodellen psychometrisch abbilden und die Antwortverhalten in bestimmten Testaufgaben mit mehreren dahinterliegenden latenten Merkmalsdimensionen in Zusammenhang bringen (W.-C. Wang & Chen, 2004). Dabei kann es zum einen so sein, dass Items eines Tests verschiedene Merkmalsdimensionen adressieren oder aber, dass zur Lösung eines Items verschiede Merkmale relevant sind. Werden bei der simultanen Messung mehrerer Kompetenzen oder Merkmalsdimensionen im Rahmen multidimensionalen adaptiven Testens (MAT) MIRT-Modellen verwendet, können Effizienzsteigerung erzielt werden und auch komplexe, mehrdimensionale theoretische Kompetenzmodelle abgebildet werden.

Das Konstrukt der ICT-Skills kann durchaus als komplex beschrieben werden, wobei verschiedene Frameworks auf unterschiedliche Weise versuchen diese Komplexität greifbar zu machen. Das im Rahmen dieser Arbeit näher betrachtete CavE-ICT-Framework zur Konstruktbeschreibung von ICT-Skills (Engelhardt et al., eingereicht) ist eines davon. Dieses Framework beschreibt vier Facetten, um Aufgaben und erforderliche Fertigkeiten und Fähigkeiten zur Lösung dieser Aufgaben zu charakterisieren und zu unterscheiden. Jede der postulierten Facetten ist in verschiedene Facettenebenen unterteilt. Im Zentrum des CavE-ICT-Frameworks steht die Facette der kognitiven Prozesse mit den fünf Ebenen „Zugreifen“, „Managen“, „Integrieren“, „Bewerten“ und „Erzeugen“, die bei der Lösung von ICT-spezifischen Aufgaben relevant sein können. Neben seiner komplexen Struktur bedingt auch die im Rahmen dieser Arbeit im Fokus

stehende verhaltensnahe Erfassung von ICT-Skills über simulationsbasierte Items Testformen zu suchen, die eine besonders effiziente und ökonomisch Messung ermöglichen. Itembearbeitungs- und Testzeiten können durch den Einsatz simulationsbasierter Items deutlich steigen (Greiff, Wüstenberg, Holt, Goldhammer & Funke, 2013), da den Testpersonen mehr Interaktions- und Explorationsmöglichkeiten geboten werden, die zudem im Sinne der eigentlichen Aufgabenlösung womöglich nicht relevant sind. Zudem ist nicht nur die zu antizipierende Bearbeitungszeit simulationsbasierter, verhaltensnaher ICT-Items hoch, sondern auch deren Entwicklungsaufwand muss als recht hoch eingeschätzt werden (Engelhardt et al., eingereicht). Um dennoch eine effiziente und ökonomische Messung von ICT-Skills zu realisieren, sollen im Rahmen dieser Arbeit Testformen vorgeschlagen werden, die gemessen am diagnostischen Erkenntnisgewinn, relativ wenig Ressourcen – wie Zeit oder Geld – beanspruchen (Schmidt-Atzert, Amelang & Fydrich, 2012).

Diese Tests sollen zudem für verschiedene Bereiche einsetzbar sein. Denn nicht nur in LSAs spielt die Messung von ICT-Skills populationsbeschreibend eine wichtige Rolle, auch individual- und/oder gruppendiagnostisch ist die Erfassung von ICT-Skills von Interesse. Gerade vor dem Hintergrund, dass Schülerinnen und Schülern möglicherweise recht heterogen in ihrer Fähigkeit mit Informations- und Kommunikationstechnologien erfolgreich umzugehen sind, ist die theoriegeleitete Erfassung von ICT-Skills umso wichtiger (Eickelmann & Bos, 2011). Um den Fähigkeitsstand in Erfahrung bringen und wissen zu können, wo man gegebenenfalls bei der Förderung ansetzen kann, ist es wichtig, einen Test zu haben, mit dem man auch auf Ebene der einzelnen Schülerinnen und Schülern Aussagen über deren ICT-Skills treffen kann.

Neben Möglichkeiten des Einsatzes multidimensionaler adaptiver Tests werden daher zur ökonomischen Erfassung von ICT-Skills im Rahmen dieser Arbeit auch ICT-Kurztests durch die gezielte Auswahl geeigneter Items zusammengestellt. Diese

Testzusammenstellung kann als ein Problem der eingeschränkten kombinatorischen Optimierung beschrieben (van der Linden, 2005) und manuell wie auch automatisiert, computerbasiert realisiert werden. Die auf diese Weise erstellten ICT-Kurztests, sollten sowohl auf Populationsebene, im Sinne einschlägiger LSAs, als auch in der Gruppen- und Individualdiagnostik einsetzbar sein. Damit widmet sich diese Arbeit der effizienten, zuverlässigen und psychometrisch adäquaten Erfassung komplexer Konstrukte konkret am Beispiel von ICT-Skills unter Einsatz computerbasierter Methoden. Dabei steht die Testentwicklung und -zusammenstellung im Fokus der Arbeit, wobei die zu entwickelnden und zu evaluierenden Messinstrumente oder Strategien im Zuge von LSAs aber auch darüber hinaus genutzt werden können.

Die ersten Kapitel dieser Arbeit versuchen die theoretische Grundlage für die später dargestellten empirischen Studien zu legen. Im Rahmen von Kapitel zwei wird in die Item-Response-Theorie eingeführt. Kapitel drei gibt anschließend einen Überblick zum computerisierten adaptiven Testen. Nach diesen eher methodisch fokussierten Kapiteln schließt in Kapitel vier eine Einführung in den zu untersuchenden Inhaltsbereichen, nämlich dem kompetenten Umgang mit ICT-Aufgaben, an. In diesem Kapitel wird das dieser Arbeit zugrundliegende Projekt und die Datenbasis vorgestellt. Kapitel fünf führt schließlich in die Fragestellungen der einzelnen empirischen Studien ein. Methode und Ergebnisse von Studie I wird in Kapitel sechs dargestellt. Ziel dieser Studie ist es bereits vor der Entwicklung von Items zur Messung von ICT-Skills Hinweise zum Umfang des zu erstellenden ICT-Itempools und zur Testlänge eines adaptiven multidimensionalen Messinstruments bereitzustellen. Kapitel sieben widmet sich Studie II, die direkt auf den Ergebnissen von Studie I aufbaut, allerdings Daten des CavE-ICT-Feldtests nutzt und darüber belastbare Informationen zur Anwendbarkeit von adaptiven Testalgorithmen zur differenzierten multidimensionalen Erfassung von ICT-Skills liefert. Schließlich erweitert Studie III den Blick von der Anwendung eines ICT-Skills Test im Rahmen von LSAs auf

die Nutzbarkeit der unter großem Aufwand entwickelten ICT-Items auch im Zuge von Individual- und Gruppendiagnostik. Hierbei wird ein Ansatz der optimierten Testzusammenstellung eines ICT-Kurztests zur linearen und fixierten Vorgabe genutzt. Die Ergebnisse jeder Studie werden direkt in den jeweiligen Kapiteln diskutiert. Kapitel neun liefert eine zusammenfassende Gesamtdiskussion, in der die Ergebnisse der einzelnen Studien zusammen und in Verbindung gebracht werden. Kernpunkte der Arbeit werden kritisch reflektiert, ein Ausblick gegeben sowie ein kurzes Fazit gezogen.

Die im Rahmen dieser Arbeit erlangten Erkenntnisse können insbesondere für die Erfassung von mehrdimensionalen Konstrukten oder facettierten Merkmalen in LSAs, aber auch darüber hinaus für individual- und gruppendiagnostische Zwecke genutzt werden.

2 Item-Response-Theorie

Modelle der *Item-Response-Theorie* (IRT) dienen dazu, von Antworten auf Testitems (engl. Item Responses) auf latente Merkmale von Personen zu schließen, die diesen Antworten zugrunde liegen. In diesem Sinne dienen IRT-Modelle der Messung latenter Merkmale und werden auch als Messmodelle oder psychometrische Modelle bezeichnet (Hartig & Goldhammer, 2010). Dabei werden die Wahrscheinlichkeiten bestimmter Antworten auf die Testitems in IRT-Modellen als Funktion des zu messenden Merkmals (Personenparameter) und den Eigenschaften der Items (Itemparameter) modelliert. Die wichtigste Itemeigenschaft ist hierbei die Itemschwierigkeit. Ein wesentliches Charakteristikum von IRT-Modellen ist, dass die Schwierigkeiten der Testitems auf einer gemeinsamen Skala mit den Merkmalsausprägungen der Personen abgebildet werden. Damit gehen auch eine Reihe IRT-spezifischer Vorteile gegenüber der *Klassischen Testtheorie* (KTT; bspw. Moosbrugger, 2012) einher. Anders als bei der IRT liegt im Rahmen der KTT der Fokus allein auf den beobachtbaren Antworten einer Person: Antwortverhalten und Merkmalsausprägung werden faktisch gleichgesetzt (Hartig & Goldhammer, 2010). Durch IRT-Modelle wird hingegen der Zusammenhang zwischen Items und latenter Merkmalsausprägung mathematisch formuliert, womit ein empirisch prüfbares Modell zur Verfügung steht. Zudem werden Schätzungen latenter Merkmale auf Basis von IRT-Modellen als unabhängig von den spezifischen Testitems angesehen, wohingegen Testwerte bei der KTT stark von den im Test vorgelegten Items abhängen. Des Weiteren wird im Rahmen der IRT angenommen, dass Itemparameter über Gruppen von Testpersonen und über wiederholte Testanwendungen hinweg invariant, also unveränderlich sind (Desjardins & Bulut, 2018). Im Gegensatz zur KTT ist im Rahmen der IRT zudem eine kriterienorientierte Interpretation von Testwerten möglich (bspw. Goldhammer & Hartig, 2012) durch die gemeinsame Skala von individuellen Merkmalen und Itemschwierigkeiten. Die Nutzung einer gemeinsamen Skala ermöglicht es auch,

unterschiedlichen Testpersonen verschiedene Items vorzulegen und dennoch Messungen der Merkmalsausprägungen dieser Personen auf der gleichen Skala vorzunehmen. Da in der KTT unterschiedliche Itemschwierigkeiten in der Testwertbildung unberücksichtigt bleiben, wäre dies im Rahmen der KTT nicht zulässig. Aus diesem Grund bietet sich die Verwendung von IRT-Modellen für computerbasiertes adaptives Testen an (CAT; siehe Kapitel 3).

In den folgenden Abschnitten (Kapitel 2.1) werden zunächst die für die vorliegende Arbeit zentralen IRT-Modelle vorgestellt. Der Fokus wird auf ein- und multidimensionale Rasch-Modelle gelegt, da im Zuge dieser Arbeit auf Daten und Ergebnisse zurückgegriffen wird (siehe Kapitel 4), die unter Verwendung des eindimensionalen Rasch-Modells erzielt wurden. Im Rahmen der Studien I und II (Kapitel 6 und 7) kommen darüber hinaus multidimensionale Rasch-Modelle zum Einsatz. Im Anschluss an die Darstellung der IRT-Modelle werden Methoden der Schätzung der Modellparameter beschrieben (Kapitel 2.2 und 2.3). Abschließend wird auf Kriterien zur Auswahl eines spezifischen Modells eingegangen (Kapitel 2.4).

2.1 IRT-Modelle

Im Rahmen der Anwendung von IRT-Modellen werden häufig die Begriffe „Skalierung“ und/oder „Kalibrierung“ gebraucht und von „IRT-skalierten“ Items oder Tests gesprochen. Dies bezieht sich auf die gemeinsame Skala für das zu messende latente Merkmal und die Schwierigkeiten der zur Messung verwendeten Items, die durch die Verwendung von IRT-Modellen hergestellt wird (Hartig & Goldhammer, 2010). Dabei werden die Parameter spezifischer IRT-Modelle (Personenparameter, Itemschwierigkeit, Itemdiskriminiation, Rateparameter) bei deren Anwendung auf empirische Daten geschätzt. Verschiede IRT-Modelle unterschieden sich dabei darin, wie der

Zusammenhang zwischen dem zu messenden Merkmal und der Antwortwahrscheinlichkeit mathematisch formalisiert wird.

Im Folgenden werden IRT-Modelle vorgestellt, die auf der Logit-Funktion basieren (zu IRT-Modellen mit anderen als logistischen Funktionen siehe Embretson & Reise, 2000). Die Logit-Funktion bewegt sich in einem Wertebereich von $0 < \text{logit}(y) < 1$. Für $y \to \infty$ nähert sie sich 1, für $y \to -\infty$ 0 an und für $y = 0$ nimmt sie den Wert 0.5 an. Sie ist wie folgt formalisiert:

$$\text{logit}(y) \equiv \frac{\exp(y)}{1+\exp(y)}. \tag{2.1}$$

Die Maßeinheit von y wird als *Logits* bezeichnet, die Skala von y als *Logit-Skala*.

Prinzipiell kann zwischen eindimensionalen und multidimensionalen IRT-Modellen unterschieden werden. Bei der Verwendung eines eindimensionalen IRT-Modells (Kapitel 2.1.1) setzt man voraus, dass nur ein latentes Merkmal die gegebenen Antworten einer Testperson auf die vorgegebenen Items determiniert. Wird allerdings angenommen, dass das gezeigte Antwortverhalten nicht nur auf die Ausprägung eines latenten Merkmals zurückzuführen ist, sondern mehrere latente Merkmale oder Merkmalsdimensionen ursächlich für das beobachtete Antwortverhalten sind, sollten multidimensionale IRT-Modelle (MIRT-Modelle) zur Anwendung gebracht werden (Kapitel 2.1.2).

2.1.1 Eindimensionale IRT-Modelle

Bei der Verwendung eindimensionaler IRT-Modelle werden zwei wesentliche Modellannahmen getroffen; diese sind Eindimensionalität und *lokale stochastische Unabhängigkeit*. Ersteres referiert darauf, dass nur ein latentes Merkmal die Bearbeitung der Testitems determiniert. Letzteres bezieht sich darauf, dass die Wahrscheinlichkeit, ein Item korrekt zu lösen, allein auf die latente Ausprägung dieses Merkmals zurückzuführen ist. Dadurch wird impliziert, dass bei Konstanthaltung der Merkmalsausprägung, die Itemantworten nur noch zufallsbedingt und folglich voneinander unabhängig sind

(Embretson & Reise, 2000). Wird das Antwortverhalten auch durch andere Merkmale bestimmt und ist somit die Annahme der lokalen stochastischen Unabhängigkeit nicht erfüllt, bietet sich die Nutzung eines multidimensionalen IRT-Modells an.

Im Folgenden werden unterschiedlich komplexe IRT-Modelle für dichotome Antwortformate vorgestellt, wobei das dichotome Rasch-Modell im Fokus steht. Dichotome IRT-Modelle, die weitere Itemparameter neben der Itemschwierigkeit berücksichtigen, werden kurz angerissen. Für die Analyse polytomer Antworten gibt es ebenfalls eine Reihe verschiedener IRT-Modelle (bspw. das Partial-Credit-Model; Masters, 1982 oder das Graded-Response-Model; Samejima, 1996), die hier allerdings nicht weiter erläutert werden.

Das Rasch-Modell. Das dichotome Rasch-Modell (Rasch, 1960) ist eines der einfachsten und zugleich am häufigsten angewendeten IRT-Modelle (Hartig & Goldhammer, 2010) und gehört zur Gruppe der Einparameter-Logistischen-Modelle (1PL-Modell). Mithilfe des Rasch-Modells lässt sich die Wahrscheinlichkeit bestimmen, mit der eine Person j ein Item i in einem Test korrekt löst, gegeben der latenten Merkmalsausprägung dieser Person θ_j und der Schwierigkeit des betrachteten Item b_i. Das dichotome Antwortformat wird über die Zufallsvariable X_{ji} abgebildet. Die Wahrscheinlichkeit ein Item zu lösen wird unter Verwendung der Logit-Funktion (siehe Formel 2.1) im Rasch-Modell wie folgt modelliert:

$$P\left(X_{ji}=1\middle|\theta_j,b_i\right)=\frac{\exp\left(\theta_j-b_i\right)}{1+\exp\left(\theta_j-b_i\right)}. \tag{2.2}$$

Dabei legt ein hoher Wert des Personenparameters eine hohe Merkmalsausprägung nahe und ein hoher Wert des Itemparameters eine hohe Schwierigkeit des entsprechenden Items. Damit ergibt sich, dass die Lösungswahrscheinlichkeit eines Items zunimmt, je höher die Merkmalsausprägung einer Person ist. Zur grafischen Darstellung des

Zusammenhangs zwischen Merkmalsausprägung und Antwortwahrscheinlichkeit werden im Kontext der IRT sogenannte itemcharakteristische Kurven (engl. Item Characteristic Curve, ICC) genutzt. Die ICC bildet die verschiedenen Lösungswahrscheinlichkeiten eines Items gegeben unterschiedlicher Merkmalsausprägungen von Personen ab. In Abbildung 2.1 werden zwei Items mit unterschiedlicher Schwierigkeit im Rasch-Modell dargestellt.

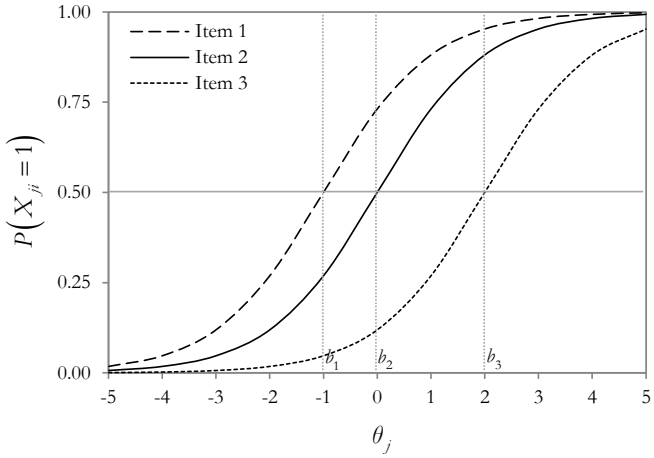

Abbildung 2.1 Itemcharakteristische Kurven (ICC) für drei Items mit Schwierigkeiten $b_1 = -1$, $b_2 = 0$ und $b_3 = 2$ im dichotomen Rasch-Modell.

Da die ICC strengmonoton steigend ist, nähert sich bei steigender Merkmalsausprägung auch die Lösungswahrscheinlichkeit eines Items dem Wert 1 an, mit abnehmender Merkmalsausprägung geht sie gegen 0. In den Randbereichen führen leicht unterschiedliche Merkmalsausprägungen daher nur zu geringen Änderungen der Lösungswahrscheinlichkeit eines Items. Im mittleren Merkmalsbereich hingegen führen leichte Änderungen der Merkmalsausprägung zu vergleichsweise starkem Anstieg oder der Abnahme der Lösungswahrscheinlichkeit. Somit ist die Steigung der ICC ein Maß dafür wie gut ein Item in bestimmten Merkmalsbereichen zwischen Personen differenzieren kann. Diese Steigung wird auch als Itemdiskrimination bezeichnet und ist im Rasch-Modell per Definition für alle Items auf 1 gesetzt. Daher ergeben sich die einzigen Unterschiede der ICC im Rasch-Modell durch den einzigen Itemparameter im Modell: die

verschiedenen Itemschwierigkeiten. In der grafischen Darstellung (siehe Abbildung 2.1) zeigen sie dies in der Verschiebung der ICC entlang der X-Achse von leichteren Items links zu schweren Items rechts.

Da für alle Items ein gleichförmiger Zusammenhang zwischen Merkmal und Antwortwahrscheinlichkeit angenommen wird, bleibt die Rangfolge der erwarteten Lösungshäufigkeiten für alle Items in allen Abschnitten der Skala bewahrt (Wilson, 2003). Dies macht die durch das Rasch-Modell definierte Skala gut interpretierbar, da es möglich ist, individuelle Testwerte durch ihre Abstände zu Itemschwierigkeiten kriteriumsorientiert zu interpretieren (Embretson, 2006; Moosbrugger, 2012).

Zwei- und Dreiprameter-Logistisches-Modell (2PL- und 3PL-Modell). Im 2PL-Modell (auch Birnbaum-Modell; Birnbaum, 1968) wird gegenüber dem Rasch-Modell einen zusätzlichen Parameter für die Itemdiskrimination a_i einbezogen. Mit ihm wird im Modell die Differenz zwischen dem Personenparameter θ_j und der Itemschwierigkeit b_i gewichtet:

$$P\left(X_{ji}=1\middle|\theta_j,a_i,b_i\right)=\frac{\exp\left(a_i\left(\theta_j-b_i\right)\right)}{1+\exp\left(a_i\left(\theta_j-b_i\right)\right)}. \tag{2.3}$$

Daraus ergibt sich, dass mit einer höheren Diskrimination a_i eines Items die Lösungswahrscheinlichkeit dieses Items stärker mit der Merkmalsausprägung der Person θ_j zusammenhängt.

Im 3PL-Modell (Birnbaum, 1968) wird dem Umstand Rechnung getragen, dass bei Items mit einem geschlossenen Antwortformat die richtige Antwort auch bei sehr niedriger Merkmalsausprägung noch geraten werden kann und sich dadurch die Lösungswahrscheinlichkeit nicht an null annähert. Mit Einführung eines Rateparameters c_i geht die Lösungswahrscheinlichkeit für ein Item i für $\theta\to-\infty$ nicht mehr gegen null,

sondern gegen c_i. Unterschiede in der Lösungswahrscheinlichkeit oberhalb des Rateparameters werden durch eine mit $(1-c_i)$ gewichteten Funktion beschrieben:

$$P\left(X_{ji}=1\middle|\theta_j,a_i,b_i,c_i\right)=c_i+(1-c_i)\frac{\exp\left(a_i\left(\theta_j-b_i\right)\right)}{1+\exp\left(a_i\left(\theta_j-b_i\right)\right)}. \tag{2.4}$$

Je höher der Rateparameter c_i, desto geringer ist die Spannbreite an Lösungswahrscheinlichkeiten, die durch Unterschiede in den Personenparametern erklärt wird. Abbildung 2.2 zeigt die ICCs für zwei Items im 2PL-Modell mit unterschiedlichen Schwierigkeiten und Diskriminationsparametern sowie im 3PL-Modell mit zusätzlich unterschiedlichen Rateparametern.

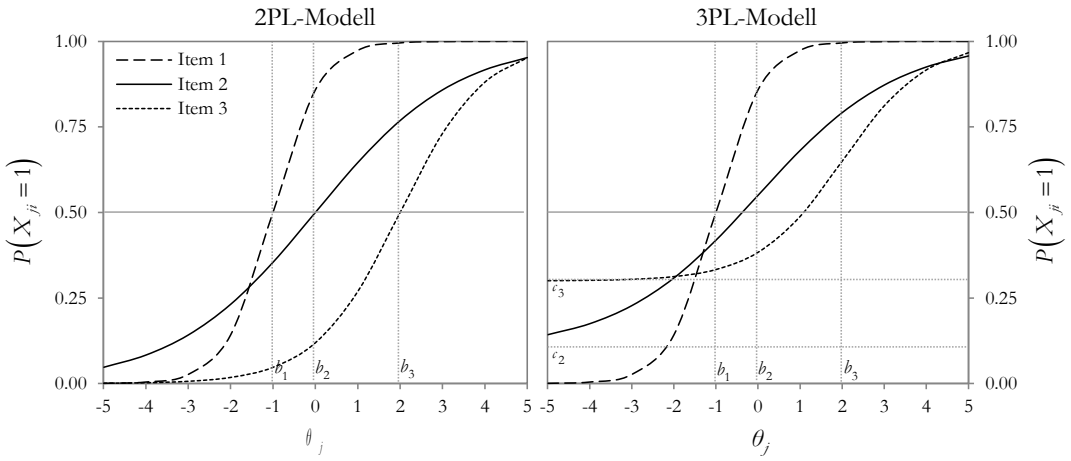

Abbildung 2.2 Itemcharakteristische Kurven (ICC) von drei Items mit Schwierigkeiten $b_1=-1$, $b_2=0$ und $b_3=2$ Diskriminationen $a_1=1.8$, $a_2=0.6$ und $a_3=1.0$ im Zweiparameter-Logistischen-Modell (2PL) links und zudem mit Rateparametern $c_1=0.0$, $c_2=0.1$ und $c_3=0.3$ im Dreiparameter-Logistischen-Modell (3PL) rechts.

Im Gegensatz zum Rasch-Modell verlaufen die ICCs nicht mehr parallel, sondern unterscheiden sich in ihrer Steigung. Damit ist die Rangfolge der erwarteten Lösungshäufigkeiten für alle Items nicht mehr in allen Abschnitten der Skala gleich. Dies

kann beispielsweise mit Blick auf den linken Teil der Abbildung 2.2 (2PL-Modell) verdeutlicht werden: Eine Person mit einer niedrigen Merkmalsausprägung von -2 hat eine höhere Wahrscheinlichkeit Item 1 zu lösen als Item 2. Dahingegen hat eine Person mit einer relativ hohen Merkmalsausprägung von 2 eine höhere Lösungswahrscheinlichkeit für Item 2 als für Item 1. Die Differenzen zwischen den Lösungswahrscheinlichkeiten von Items verändern sich also in Abhängigkeit von der Merkmalsausprägung (Moosbrugger, 2012). Damit verfügt das Rasch-Modell im Vergleich zu mehrparametrigen Modellen über Vorteile im Hinblick auf die kriteriumsorientierte Testwertinterpretationen.

Item- und Testinformationsfunktion. Die Iteminformation ist ein zentrales Konzept der IRT zur Beschreibung von Items und bezieht sich darauf, dass jedes von einer Person bearbeitete Item psychometrische oder diagnostische Information zur Schätzung der Merkmalsausprägung der Person beiträgt. Die Höhe der Iteminformation eines Items i in Abhängigkeit von θ wird in der *Iteminformationsfunktion* (IIF) $I_i\left(\theta_j\right)$ ausgedrückt. Darin fließt die Wahrscheinlichkeit $p_i\left(\theta\right)$, mit der eine Person mit der latenten Merkmalsausprägung θ ein Item i korrekt löst, und die entsprechende Gegenwahrscheinlichkeit $q_i\left(\theta\right)$ ein. Für das Rasch-Modell ergibt sich:

$$I_i\left(\theta_j\right) = p_i\left(\theta_j\right) q_i\left(\theta_j\right). \tag{2.5}$$

IIF von Rasch-skalierten Items haben die gleiche Form und sind lediglich auf der θ -Skala je nach Itemschwierigkeit unterschiedlich lokalisiert. Dabei wird die Messgenauigkeit eines Items maßgeblich davon beeinflusst, inwieweit Itemschwierigkeit und Personenparameter übereinstimmen. Für das 2PL-Modell wird die in Formel 2.5 dargestellte IIF mit der Diskrimination gewichtet, was dazu führt, dass die IIF schmaler und steiler oder breiter und flacher verlaufen können. Für das 3PL-Modell fließt zudem noch der Rateparameter in die IIF mit ein.

IIF von Items, die auf einer gemeinsamen θ-Skala abgebildet werden, können zur Testinformation aufaddiert werden. Die *Testinformationsfunktion* (TIF) $I(\theta_j)$ gibt wieder wie gut ein aus N Items gebildeter Test zur Messung der Merkmalsausprägung θ geeignet ist:

$$I(\theta_j) = \sum_{i=1}^{N} I_i(\theta). \tag{2.6}$$

Aus der IIF wie auch der TIF kann der Standardmessfehler (*SE*) der Schätzung der individuellen Merkmalsausprägung θ einer Person j berechnet werden, der eine Einschätzung der Messpräzision eines Items oder Tests liefert. Der Zusammenhang zwischen Standardmessfehler und Iteminformation lässt sich wie folgt beschreiben:

$$SE(\theta_j) = \frac{1}{\sqrt{I_i(\theta_j)}}, \tag{2.7}$$

wobei $I_i(\theta_j)$ durch $I(\theta_j)$ ersetzt werden kann, wenn nicht nur die Messpräzision eines Items sondern die eines aus mehreren Items gebildeten Tests eingeschätzt werden soll. Aus der Formel 2.7 wird deutlich, dass die Messgenauigkeit in Abhängigkeit von der Merkmalsausprägung θ bestimmt wird. Entsprechend lässt sich so die Eignung eines Items oder Tests für bestimmte Bereiche der Merkmalsausprägungen beurteilen. Abbildung 2.3 stellt den Zusammenhang zwischen TIF und *SE* grafisch dar. Man erkennt, dass am höchsten Punkt der TIF, wo die meiste diagnostische Information vorliegt, die Messgenauigkeit am höchsten ist. Wohingegen extrem hohe wie niedrige Merkmalsausprägungen nur ungenau geschätzt werden können, da weniger Items mit extremen Schwierigkeiten im Test sind.

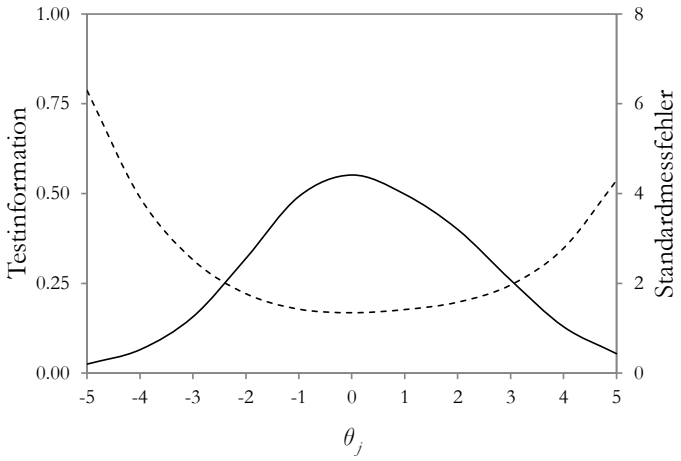

Abbildung 2.3 Zusammenhang von Testinformationsfunktion (TIF; durchgezogene Linie) und Standardmessfehler (SE; unterbrochene Linie) für die drei in Abbildung 2.1 dargestellten Items

Die Iteminformation ist von hoher praktischer Relevanz im Rahmen des CAT (siehe Kapitel 3), da dabei im Testverlauf entsprechend der aktuell geschätzten Merkmalsausprägung einer Person gezielt Items ausgewählt werden, die maximale diagnostische Information liefern.

2.1.2 Multidimensionale IRT-Modelle

Mitunter bieten die zuvor dargestellten unidimensionalen IRT-Modelle keine ausreichende Grundlage zur Modellierung von Testergebnissen. Dabei kann es zum einen so sein, dass Items eines Tests verschiedene Merkmalsdimensionen adressieren oder aber, dass zur Lösung eines Items verschiede Merkmale relevant sind. Beide Überlegungen zeigen wie latente Personenmerkmale auf unterschiedliche Weise interagieren können, um ein bestimmtes Antwortverhalten zu zeitigen. Multidimensionale IRT-Modelle (MIRT-Modelle) wurden daher unter anderem vorgeschlagen, um der zunehmenden Komplexität moderner Messungen gerecht zu werden (W.-C. Wang & Chen, 2004). MIRT-Modelle kommen also zum Einsatz, wenn das Antwortverhalten von Personen auf eine Menge von Items nicht nur auf eine latente Merkmalsdimension θ, sondern auf D latente

Merkmalsdimensionen $\boldsymbol{\theta} = (\theta_1, \theta_2, ... \theta_D)$ zurückgeführt werden soll (bspw. Reckase, 2009). Unterscheidungen von MIRT-Modellen beziehen sich zum einen auf deren Anwendung (konfirmatorisch oder explorativ), zum anderen auf deren Komplexität, in Form der Anzahl der zur Lösung eines Items angenommenen Merkmalsdimensionen (Between- und Within-Item-MIRT-Modelle). Im Fall von Within-Item-Modellen können zudem unterschiedliche Annahmen bezüglich der Relation der beteiligten Merkmalsdimensionen zueinander getroffen werden (kompensatorische und nicht-kompensatorische Modelle).

Im Rahmen der vorliegenden Arbeit wird konfirmatorisch vorgegangen, indem Modellparameter für theoretisch angenommene Merkmalsdimensionen geschätzt werden. Der Zusammenhang zwischen den Items und den angenommenen Dimensionen ist vorab spezifiziert. Liegen solche theoretisch fundierten Annahmen über die Beziehung von Items und Merkmalsdimensionen nicht vor, gilt es, diese aus den empirischen Daten abzuleiten, was einem explorativen Vorgehen entspricht (Embretson & Reise, 2000; Desjardins & Bulut, 2018). Zudem werden im Zuge dieser Arbeit ausschließlich Between-Item-MIRT-Modelle angewendet. Das heißt, jedes latente Merkmal beziehungsweise jede latente Merkmalsdimension wird ausschließlich durch eine definierte Menge von Items gemessen. Damit ist ein Item nur einem durch den Test zu messenden Merkmal beziehungsweise einer Merkmalsdimension zugeordnet. Hingegen kann in einem Within-Item-MIRT-Modell jedes Item auch mehreren durch den Test zu messenden latenten Merkmalsdimensionen zugeordnet sein. Within-Item-Modelle, die komplex Teststrukturen abbilden, können nochmals durch die Art und Weise des Zusammenwirkens mehrerer Dimensionen beim Zustandekommen einer Itemantwort unterschieden werden. Kompensatorischen Modellen liegt die Annahme zugrunde, dass Schwächen in einem latenten Merkmal beziehungsweise einer Merkmalsdimension durch Stärken in einer anderen latenten Merkmalsdimension ausgeglichen werden können. Für nicht-kompensatorische Modelle wird angenommen, dass die Schwächen in einer

Merkmalsdimension nicht durch Stärken in einer anderen kompensiert werden können. Eine umfassende Diskussion multidimensionaler IRT-Modelle findet sich beispielsweise in Reckase (2009).

Multidimensionale Erweiterung des Rasch-Modells. Adams, Wilson und Wang (1997) führten eine multidimensionale Erweiterung des Unidimensional-Random-Coefficients-Multinomial-Logit-Model (RCML-Modell; Adams & Wilson, 1996) ein, welches eine Generalisierung des Rasch-Modells darstellt und damit eine ganze Bandbreite gängiger Modelle, die auf dem Rasch-Modell basieren, integriert (bspw. das Partial Credit Model; Masters, 1982 oder das Multifacetten-Rasch-Modell; Linacre, 1994). Die multidimensionale Erweiterung des RCML-Modells zum Multidimensional-Random-Coefficients-Multinomial-Logit-Model (MRCML-Modell) bietet ebenso eine allgemeine Formulierung und kann daher ebenfalls nicht nur mit dichotomen, sondern auch mit polytomen Antwortformaten umgehen. Die zu messenden latenten Merkmale definieren einen D-dimensionalen Raum, in dem die Positionierung einer Person j durch den Merkmalsvektor $\boldsymbol{\theta}_j = \left(\theta_{j_1}, \theta_{j_2}, ..., \theta_{j_D}\right)$ repräsentiert wird. Der Notation von Adams et al. (1997) folgend ist die Wahrscheinlichkeit einer Person j, eine Antwortkategorie k, wobei $k = 0, 1, 2, \ldots, K$, eines Items i in einem D-dimensionalen Test zu wählen

$$P\left(X_{ijk} = 1 \middle| \boldsymbol{\xi}, \boldsymbol{\theta}_j\right) = \frac{\exp\left(\mathbf{b}'_{ik}\boldsymbol{\theta}_j + \mathbf{a}'_{ik}\boldsymbol{\xi}\right)}{\sum_{k=0}^{K} \exp\left(\mathbf{b}'_{ik}\boldsymbol{\theta}_j + \mathbf{a}'_{ik}\boldsymbol{\xi}\right)}. \tag{2.8}$$

Dabei ist $\mathbf{b}_{ik}$ ein Score-Vektor der Antwortkategorie k eines Items i über die D Merkmalsdimensionen, sodass $\mathbf{b}_{ik} = \left(b_{ik_1}, b_{ik_2}, ..., b_{ik_D}\right)$. $\boldsymbol{\xi}$ ist ein Vektor der Itemparameter. $\mathbf{a}_{ik}$ ist ein Design-Vektor der Antwortkategorie e eines Items i über die D Merkmalsdimensionen. Die Vektoren $\mathbf{b}_{ik}$ und $\mathbf{a}_{ik}$ beziehen sich nicht direkt auf die

Itemschwierigkeit oder Itemdiskrimination, vielmehr handelt es sich um Gewichte, die sich aus der Teststruktur ableiten (Desjardins & Bulut, 2018).

Die Korrelation zwischen den latenten Merkmalen oder Merkmalsdimensionen dient als zusätzliche Informationsquelle bei der Schätzung der Modellparameter. Dies trägt substantiell zur Verbesserung der Messpräzision bei, vor allem wenn der Test kurz und die Anzahl der zu schätzenden Merkmalsdimensionen hoch ist (W.-C. Wang & Chen, 2004). Das MRCML-Modell kann im Rahmen von Tests mit Between- aber auch mit Within-Item-Multidimensionalität angewendet werden.

2.2 Personenparameterschätzung

Das Ziel einer IRT-basierten Diagnostik besteht darin, für eine Person j auf der Grundlage ihres Antwortmusters einen individuellen Personenparameter θ_j zu schätzen. Zwei Ansätze zur Personenparameterschätzung werden im Folgenden vorgestellt zum einen Maximum-Likelihood-Schätzung (ML-Ansatz) und zum anderen Bayesianische Schätzung (Reckase, 2009; Segall, 1996). Bei den folgend vorgestellten Schätzverfahren werden die Itemparameter der bearbeiteten Items als bereits bekannt angenommen.

2.2.1 Maximum-Likelihood

Bei der Bestimmung des Personenparameters durch Maximum-Likelihood-Schätzung (engl. Maximum-Likelihood-Estimation; MLE) schließt man vom beobachteten Antwortmuster auf diejenige Ausprägung von θ, welche die *Likelihood* (die Plausibilität) des Antwortmusters maximiert (bspw. Embretson & Reise, 2000; Hambleton & Swaminathan, 2010). Unter der für IRT-Modelle grundlegenden Annahme der lokalen stochastischen Unabhängigkeit kann die Likelihood-Funktion L für ein Antwortmuster $x_{j_1},x_{j_2},...,x_{j_n}$ als das Produkt der Likelihoods der einzelnen Itemantworten für N Items bestimmt werden:

$$L\left(x\middle|\theta_j\right) = L\left(x_{j_1}, x_{j_2}, ..., x_{j_n}\middle|\theta_j\right) = \prod_{i=1}^{N} P_i\left(\theta_j\right)^{x_{ji}} Q_i\left(\theta_j\right)^{1-x_{ji}} . \tag{2.9}$$

Die Exponenten x_{ji} und $1-x_{ji}$ haben zur Folge, dass bei $x_{ji}=1$ die Likelihood einer richtigen Antwort, bei $x_{ji}=0$ die Likelihood einer falschen Antwort in die Likelihood des Antwortmusters eingeht (Hartig & Goldhammer, 2010). Gegeben des angenommenen IRT-Modells definiert $P_i\left(\theta_j\right)$ die Wahrscheinlichkeit, ein Item i zu lösen. $Q_i\left(\theta\right)=1-P_i\left(\theta\right)$ entspricht der Wahrscheinlichkeit, das Item nicht zu lösen.

Der MLE-Schätzer gilt allerdings vor allem bei kurzen Tests als verzerrt. Alternativ kann der gewichtete MLE-Schätzer (engl. Weighted-Likelihood-Estimation oder auch Warm's-Likelihood-Estimation, WLE; Warm, 1989) verwendet werden. Dabei wird die Likelihood-Funktion mit der Wurzel der Testinformationsfunktion gewichtet und dadurch eine geringere mittlere Abweichung der Schätzwerte vom Erwartungswert im Vergleich zum MLE erzielt (Warm, 1989). Zudem lassen sich durch WLE-Schätzer auch endliche Werte für die Personen berechnen, die kein Item richtig oder alle Items richtig beantwortet haben (invariantes Antwortmuster), was bei der Verwendung der MLE-Schätzung nicht möglich ist. Eine multidimensionale Erweiterung des WLE-Schätzers wurde von C. Wang (2015) vorgeschlagen.

2.2.2 Bayesianische Ansätze

Die folgend dargestellten Bayesianischen Ansätze bauen direkt auf dem ML-Ansatz auf. Es wird allerdings nicht die Plausibilität des beobachteten Antwortmusters unter der Bedingung des Personenparameters θ betrachtet, sondern die Wahrscheinlichkeit des Personenparameters unter der Bedingung des beobachteten Antwortmusters (Rost, 2004). Bayesianische Schätzer für Personenparameter zeichnen sich dadurch aus, dass nicht nur das Antwortmuster der jeweiligen Person für deren Merkmalsschätzung verwendet wird, sondern zusätzliche Informationen – in Form einer A-priori-Verteilung $f\left(\theta_j\right)$ der

Merkmalsausprägungen in der Population – in die Schätzung mit einfließen. Oft wird für die A-priori-Verteilung eine Standardnormalverteilung angenommen, wobei jede beliebige Verteilung genutzt werden könnte (Hartig & Goldhammer, 2010).

Gängige Methoden zur Bestimmung von θ bei der Bayesianischen Schätzung sind die Bestimmung des Maximums der A-posteriori-Verteilung (engl. Maximum-A-posteriori-Estimation, MAP; Bock & Aitkin, 1981) oder die Bestimmung des Mittelwerts der A-posteriori-Verteilung (engl. Expected-A-posteriori-Estimation, EAP; Bock & Mislevy, 1982). Zur Berechnung des MAP-Schätzers wird die A-priori-Verteilung $f\left(\theta_j\right)$ mit der Likelihood-Funktion $L\left(x\middle|\theta_j\right)$ multipliziert. Die resultierende Verteilung ist proportional zur A-posteriori-Verteilung $f\left(\theta_j\middle|x_{j_1},x_{j_2},...,x_{j_n}\right)$, welche die Wahrscheinlichkeit der Merkmalsausprägung θ unter der Bedingung des Antwortmusters angibt:

$$f\left(\theta_j\middle|x_{j_1},x_{j_2},...,x_{j_n}\right) \propto L\left(x_{j_1},x_{j_2},...,x_{j_n}\middle|\theta_j\right)\cdot f\left(\theta_j\right). \tag{2.10}$$

Wie der Name schon sagt, wird beim MAP-Ansatz diejenige θ-Ausprägung gesucht, welche die A-posteriori-Verteilung maximiert. Im Zuge der EAP-Schätzung wird dagegen der Erwartungswert der A-posteriori-Verteilung zur Schätzung des Personenparameters θ verwendet.

Wie der WLE-Schätzer erlaubt auch die Bayesianische Schätzung eine Bestimmung des Personenparameters für Personen mit invariantem Antwortverhalten. Ein weiterer Vorteil ist, dass die Personenparameterschätzung präziser vorgenommen werden kann, wenn zusätzliche Informationen zur Verteilung der Personenparameter in der Population berücksichtigt werden (bspw. Embretson & Reise, 2000). Dadurch können Bayesianische Schätzer auch bei kurzen Tests angewendet werden, sind dann aber eher in Richtung des Maximums beziehungsweise des Erwartungswerts der A-priori-Verteilung verzerrt (Lord,

1986). Mit zunehmender Testlänge nimmt dieser Effekt ab. Zudem muss festgehalten werden, dass Bayesianische Schätzungen beträchtlich verzerrt sein können, wenn falsche Annahmen über die A-priori-Verteilung getroffen werden (Embretson & Reise, 2000).

2.3 Itemkalibrierung

In den vorangegangenen Abschnitten wurde davon ausgegangen, dass die für die Personenparameterschätzung erforderlichen Itemparameter bekannt sind. Bei der Entwicklung eines neuen Tests liegen zunächst allerdings weder Personen- noch Itemparameter vor, sodass beide gemeinsam geschätzt werden müssen. Das primäre Ziel einer Kalibrierungsstudie liegt in der Bestimmung der Itemparameter und der Prüfung der Itemgüte. Auf Basis dieser Analysen werden Items für den skalierten Itempool oder Test selektiert oder davon eliminiert. Daher ist die Kalibrierungsstudie ein integraler Bestandteil der Entwicklung eines IRT-basierten Tests. Die Testbedingungen in der Kalibrierungsstudie sollten stark mit denen der später geplanten Testanwendung übereinstimmen, um Vergleichbarkeit zu gewährleisten. Des Weiteren ist im Rahmen der Kalibrierungsstudie sicherzustellen, dass alle Testpersonen tatsächlich über eine Ausprägung des zu messenden Merkmals verfügen und dass diese Merkmalsausprägungen über den gesamten Merkmalsbereich streuen, der auch mit dem späteren Test abgebildet werden soll (Frey, in Druck). Die erforderliche Größe der Kalibrierungsstichprobe richtet sich maßgeblich nach der Anzahl der zu kalibrierenden Items und dem genutzten IRT-Modell. Grob gesagt werden Antworten von mehr Testpersonen benötigt, je mehr Items kalibriert werden sollen und je mehr Parameter im genutzten Modell zu schätzen sind.

Bei der Kalibrierung können auch sogenannte balancierte unvollständige Block-Designs (bspw. Frey, Hartig & Rupp, 2009) zum Einsatz kommen. Dies ist vor allem dann sinnvoll, wenn größere Itempools kalibriert werden und nicht jeder Testperson in einer angemessenen Zeit alle Items vorgelegt werden können. Bei unvollständigen

Block-Designs wird jeder Testperson nur eine Teilmenge der zu kalibrierenden Items vorgelegt, zugleich können potentielle Störgrößen wie die Vorgabehäufigkeit von Items und Itemposition während der Testungen aber ausbalanciert werden. Die Anzahl der Antworten pro Item und die Größe der Stichprobe sind dann entsprechend nicht mehr identisch. Als Richtlinien für eine zuverlässige Schätzung von Itemparametern in eindimensionalen IRT-Modellen werden für das Rasch-Modell mindestens 200 Antworten pro Item, für das 2PL-Modell 500 und für das 3PL-Modell 1.000 Antworten pro Item vorgeschlagen (Eggen, 2008 oder de Ayala, 2009). Für die Kalibrierung von Itempools, die im Rahmen von CAT eingesetzt werden sollen, sind diese Richtlinien als Untergrenzen zu verstehen (Frey, in Druck).

Für die Schätzung der Itemparameter bei gleichzeitig unbekannten Personenparametern stehen unterschiedliche Maximum-Likelihood-Verfahren im Rahmen der IRT zur Verfügung (bspw. Joint-Maximum-Likelihood oder Marginal-Maximum-Likelihood; Embretson & Reise, 2000). Die MML-Schätzung ist eine häufig verwendete Technik bei der Itemkalibrierung, die bei eindimensionalen ein- und mehrparametrigen Modellen wie auch bei MIRT-Modellen angewendet werden kann (bspw. Glas, 2010). Zudem können bei der Schätzung der Itemparameter auch Posteriori-Verteilungen genutzt werden (Bayes-Modal-Estimates; Wainer & Mislevy, 2000).

Bei der Analyse der psychometrischen Güte von Items und zur Entscheidungsfindung hinsichtlich der Selektion von Items für einen IRT-skalierten Itempool oder Test werden meist verschiedene Kriterien herangezogen. Die Prüfung des *Itemfit* bezieht sich darauf, ob die durch das gewählte IRT-Modell und die geschätzten Itemparameter vorhergesagten Antworthäufigkeiten in Abhängigkeit von der geschätzten Merkmalsausprägung mit den beobachteten Antworthäufigkeiten übereinstimmen (Embretson & Reise, 2000). Ein häufig verwendetes Residuen-basiertes Fit-Maß ist die Mean-Squared-Fit-Statistic (MNSQ) oder der gewichtete MNSQ (engl. Weighted MNSQ;

WMNSQ), der auf einem standardisierten Vergleich zwischen erwarteten und beobachteten Werten beruht (Wu, Adams, Wilson & Haldane, 2007). Der Erwartungswert des MNSQ wie des WMNSQ für das eindimensionale Rasch-Modell liegt bei 1. Zusätzlich kann der Itemfit inferenzstatistisch überprüft werden. Die entsprechenden t-Werte des MNSQ oder WMNSQ sollten dann zwischen -2 und 2 liegen (Bond & Fox, 2007).

Des Weiteren wird häufig die Itemdiskrimination zur Prüfung der Güte von Items herangezogen. Dazu kann die punkt-biseriale Korrelation von Itemtestwert und Gesamttestwert herangezogen werden, um einzuschätzen, wie gut Items zwischen Personen mit unterschiedlichen Personenparametern differenzieren. Diese Korrelation entspricht dann einer Itemdiskrimination, wie sie in der KTT verwendet wird. Oftmals werden zudem differentielle Itemeffekte (engl. Differential-Item-Functioning, DIF; Osterlind & Everson, 2009) untersucht, da diese dafür sprechen können, dass ein Test nicht gegenüber allen Personengruppen fair ist (Zumbo, 1999). Wenn Personen unterschiedlicher Gruppen (bspw. unterschiedlichen Geschlechts) mit derselben latenten Merkmalsausprägung nicht die gleiche Wahrscheinlichkeit haben, ein Item korrekt zu beantworten, spricht man von DIF im Sinne eines systematischen Effekts. Die statistische Analyse von DIF beim eindimensionalen Rasch-Modell kann über die Anwendung eines Multifacetten-Rasch-Modells (Linacre, 1994) erfolgen (Wu et al., 2007).

2.4 Modellangemessenheit

Für die Frage der Angemessenheit eines verwendeten IRT-Modells sind einerseits inhaltliche Anforderungen (theoriegeleitete Beurteilung der Angemessenheit) und andererseits statistische Kriterien (datengeleitete Beurteilung der Angemessenheit) in Betracht zu ziehen. Wie gut die erhobenen Daten das gewählte IRT-Modell abbilden, wird über den sogenannten Modellfit statistisch untersucht (bspw. Embretson & Reise, 2000; Hambleton & Swaminathan, 2010). Häufig werden verschiedener IRT-Modelle

angewendet und mithilfe statistischer Kriterien deren relative Modellpassung zu den gegebenen Daten beurteilt. Zum Beispiel können durch den Likelihoodquotienten-Tests (engl. Likelihood Ratio Test) die Deviance-Statistiken zweier verschachtelter (d. h. in einander überführbarer) Modelle untersucht werden. Die Deviance gibt dabei an, wie weit das verwendete Modell und die erhobenen Daten voneinander abweichen und ist somit ein Maß für die mangelnde Passung. Informationstheoretische Kriterien zur Beurteilung des Modelfit basieren wie der Likelihoodquotienten-Test auf der Plausibilität der zu vergleichenden Modelle, wobei die Anzahl der Parameter und der Wert der Log-Likelihood der Modelle verrechnet werden. Häufig genutzte informationstheoretische Kriterien sind Akaike's-Information-Criterion (AIC; Akaike, 1973) und Bayes-Information-Criterion (BIC; Schwarz, 1978). Beide können auch angewendet werden, wenn die zu vergleichenden Modelle nicht verschachtelt sind. Der AIC ist allerdings bei geringer Itemanzahl und hohen Antwortmusterhäufigkeiten, der BIC bei größeren Itemanzahlen und wenigen Antwortmusterhäufigkeiten vorzuziehen (Rost, 2004).

Durch die Hinzunahme von weiteren Parametern zusätzlich zum Schwierigkeitsparameter, wie bei 2PL- und 3PL-Modellen, erreicht man aufgrund der höheren Flexibilität in der Modellierung in der Regel Verbesserungen der Modellpassung. Allerdings sollte eine Überparametrisierung vermieden und sparsamere Modelle bevorzugt werden, sofern die Annahmen dieser Modelle nicht verletzt werden. Deshalb sind, sofern theoretische Annahmen nicht dagegen sprechen, bei gleicher Passung auch eindimensionale Modelle den mehrdimensionalen Modellen vorzuziehen, da sie weniger Annahmen über das Antwortverhalten machen (Thompson & Weiss, 2011; Wise & Kingsbury, 2000). Die Anzahl und inhaltliche Definition der Dimensionen kann entweder theoriebasiert sein (konfirmatorische Modellierung) oder aus dem Vergleich von Modellen mit unterschiedlicher Anzahl von Dimensionen und der Auswahl des aufgrund statistischer Kriterien am besten passenden Modells (explorative Modellierung) bestimmt

werden. Im Hinblick auf die Anzahl und inhaltliche Definition der im Modell vorgesehenen Dimensionen wurde von de Ayala (2009) angemerkt, dass die angenommenen Dimensionen nützlich und psychologisch bedeutsam sein müssen, was letztlich eine Validitätsfrage darstellt. Allgemein ist die Verwendung des sparsamsten Modells, welches dennoch angemessen die Antworten der Probanden abbildet, zu empfehlen.

In diesem Kapitel wurden die Grundidee sowie einige Modelle der IRT vorgestellt. IRT-skalierte Items zur Erfassung des komplexen Konstrukts ICT-Skills (siehe Kapitel 4) werden im Rahmen dieser Arbeit genutzt, um in den Studien I und II (Kapitel 6 und 7) unterschiedliche multidimensionale adaptive Algorithmen zu erproben. In Studie III (Kapitel 8) werden ICT-Items zu Kurztests zusammengestellt und evaluiert. In allen Studien kommen IRT-Modelle zur Schätzung der Personenparameter zum Einsatz.

3 Computerisiertes adaptives Testen

Viele bekannte und häufig verwendete Testverfahren bestehen aus einem festen Set an Items, die allen zu testenden Personen in einer festgesetzten Reihenfolge vorgelegt werden. Mit der zunehmenden Nutzung computergestützter Testverfahren ergeben sich neue Möglichkeiten, Tests stärker auf die Merkmale und Bedürfnisse der Befragten abzustimmen. *Computerisiertes adaptives Testen* (CAT) ist eine solche Möglichkeit, die zunehmend Aufmerksamkeit und Anwendung findet. Bei der Anwendung adaptiver Testverfahren werden einer Person Items vorgelegt, die – entsprechend des gezeigten Antwortverhaltens – die meiste diagnostische Information hinsichtlich einer zu messenden Merkmalsausprägung liefern. Dabei wird zunächst ein Item administriert und die vom Befragten gegebene Antwort bewertet. Darauf basierend wird die Merkmalsausprägung (bspw. Fähigkeit, Wissen oder auch Einstellung) der Person geschätzt und darauf basierend ein neues Item ausgewählt und vorgelegt. Damit erhält jede Person einen an ihr individuelles Antwortverhalten angepassten Test, der sozusagen entsprechend ihrer individuellen Merkmalsausprägung maßgeschneidert ist (Frey, 2012). Daraus ergibt sich, dass beim adaptiven Testen Personen mit einer hohen Merkmalsausprägung eher schwierigere Items im Testverlauf vorgelegt werden als Personen mit einer niedrigeren Merkmalsausprägung. Zudem werden weniger, idealerweise sogar gar keine Items vorgelegt, die für die Testperson deutlich zu leicht oder zu schwer sind und somit immer beziehungsweise nie gelöst werden können. Durch die Vermeidung der Vorgabe von solchen Items, die wenig diagnostische Information liefern, ergibt sich beim adaptiven Testen im Vergleich zum linearen, nicht-adaptiven Testen typischerweise eine deutlich geringe Anzahl vorzugebender Items bei vergleichbarer *Messpräzision* und somit eine Steigerung der *Messeffizienz*. Die Messpräzision beschreibt die Messgenauigkeit eines Testwertes und kann als durchschnittliche oder erwartete quadrierte Abweichung zwischen geschätzten und wahren Personenparametern gebildet werden

(Segall, 2005). Somit kann der Standardfehler der Personenparameterschätzung zur Einschätzung der Messpräzision herangezogen werden. Aus dem Verhältnis von Messpräzision zu Testlänge ergibt sich die Messeffizienz (Frey, 2012), was bedeutet, dass bei hoher Messpräzision und kurzer Länge ein Test als besonders messeffizient einzuschätzen ist.

Die Grundlage für CAT bieten IRT-Modelle (siehe Kapitel 2); diese werden als psychometrische Modelle genutzt und CAT-Itempools IRT-skaliert. Die IRT-Skalierung ermöglicht die Nutzung einer gemeinsamen Metrik für Itemschwierigkeiten und individuelle Merkmalsausprägungen. Dadurch kann zum einen der diagnostische Informationsgehalt von Items für die zu testende Person bestimmt werden und im Testverlauf das jeweils informativste Item zur Administration ausgewählt werden. Zum anderen erlaubt sie eine Vergleichbarkeit von Testwerten verschiedener Personen, obwohl ihnen im CAT unterschiedliche Items vorgelegt werden.

Im Folgenden wird zunächst auf den typischen Ablauf eines CAT eingegangen und die einzelnen Schritte und bestimmenden Elemente näher erläutert (Kapitel 3.1). Da die Mehrzahl der bislang genutzten CAT auf eindimensionalen IRT-Modellen basieren, wird sich in diesem ersten Teil auf eindimensionale CAT bezogen. Im Anschluss daran wird auf mehrdimensionale adaptive Tests eingegangen (Kapitel 3.2). Zudem wird im Folgenden der Fokus auf voll adaptive computerisierte Tests gelegt, bei denen tatsächlich jedes Item adaptiv administriert wird. Speziellere Formen wie CAT mit Itemgruppen, die sich auf einen gemeinsamen Stimulus beziehen (sog. Testlets) oder mehrstufige Tests (engl. Multistage Tests) nicht vorgestellt (Näheres zu diesen Testformen findet sich bpsw. in van der Linden & Glas, 2010).

3.1 Ablauf und bestimmende Elemente eines CAT

In Abbildung 3.1 werden das Grundprinzip und die bestimmenden Elemente von CAT mit Hilfe eines Flussdiagramms veranschaulicht.

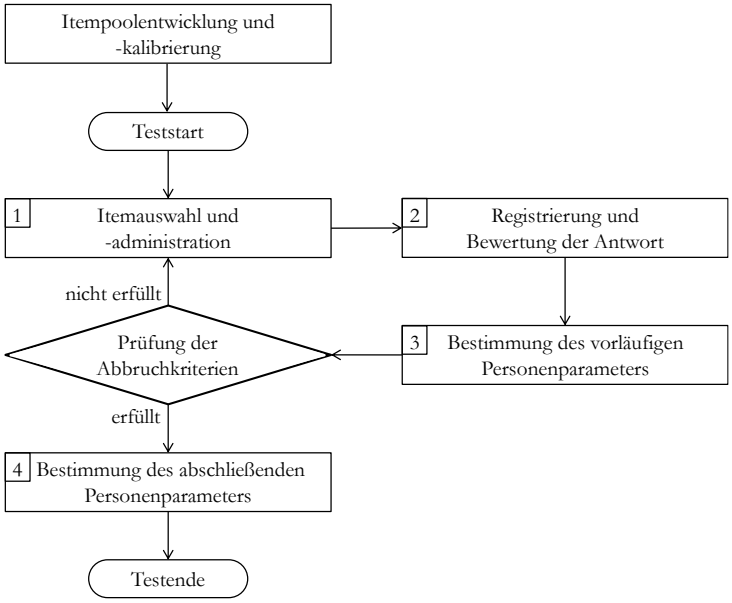

Abbildung 3.1 Flussdiagramm zum typischen Ablauf computerisierter adaptiver Tests (in Anlehnung an Thompson & Weiss, 2011).

Ein CAT startet mit der Auswahl eines ersten Items aus dem zur Verfügung stehenden IRT-kalibrierten Itempool und dessen Administration (1). Die Antwort der Testperson wird registriert und bewertet (2). Damit liegt empirisch gewonnene Information vor, auf deren Basis eine vorläufige Schätzung der individuellen Merkmalsausprägung der Testperson vorgenommen werden kann (3). Dieser vorläufig geschätzte Personenparameter kann dann wiederum genutzt werden, um ein neues Item aus dem Pool zur Vorlage auszuwählen, welches die größte diagnostische Information zu liefern verspricht. Für die Itemauswahl wird ein sogenanntes statistisches Optimalitätskriterium herangezogen, welches grundsätzlich das Ziel hat, die erwartete Varianz der vorläufigen Merkmalsschätzung $\hat{\theta}$ zu reduzieren (Frey & Seitz, 2009); das

heißt, mit jeder Itemauswahl die Schätzung der Merkmalsausprägung zu präzisieren. Bei vielen CAT werden zudem weitere, nicht-statistische Nebenbedingungen an die Testzusammenstellung während der Itemauswahl berücksichtigt. Die Schritte 1 bis 3 wiederholen sich solange, bis die vor Testbeginn definierten Abbruchkriterien erfüllt sind. Zum Testende wird eine abschließende Schätzung der zu messenden individuellen Merkmalsausprägung vorgenommen (4). Das Regelsystem, das spezifiziert, wie die Itemauswahl zu Beginn und im Verlauf des CAT erfolgt, wie mit nicht-statistischen Nebenbedingungen an die Testzusammenstellung umgegangen wird, wie die individuelle Merkmalsausprägung von Testpersonen geschätzt wird und welche Kriterien zur Beendigung des CAT erfüllt sein müssen, wird oftmals als *adaptiver Algorithmus* bezeichnet (Eggen, 2008; Frey, 2012).

Abbildung 3.2 stellt den für CAT zentralen Aspekt der Itemauswahl im Testverlauf für einen eindimensionalen Test auf Basis eines 1PL-Modells grafisch dar: Der dargestellten Testperson mit einer wahren individuellen Merkmalsausprägung von 1 wird zu Testbeginn ein eher leichtes Item vorgelegt. Im Verlauf der Testung nähern sich die Schwierigkeiten der ausgewählten Items dem Personenparameter an.

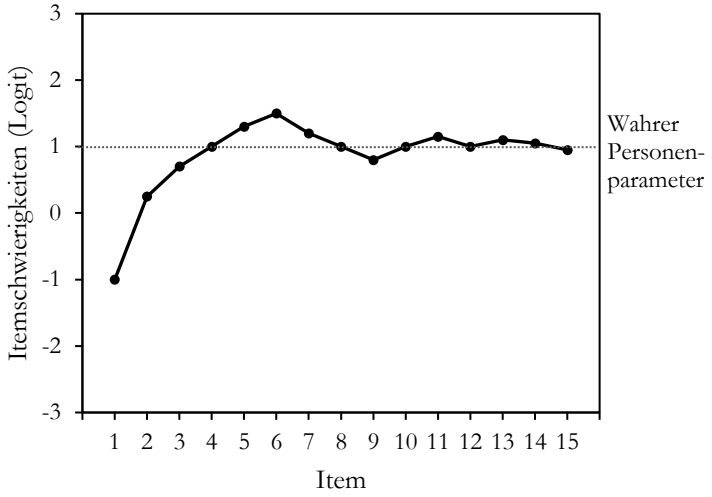

Abbildung 3.2 Darstellung zum Ablauf einer adaptiven Testung (in Anlehnung an Frey, 2012).

3.1.1 Itempool

Die Voraussetzung für CAT ist, dass alle zur Verfügung stehenden Testitems IRT-kalibriert wurden. Daher ist eine Kalibrierungsstudie ein wesentlicher Bestandteil der Entwicklung eines CAT im Vorfeld der CAT-Anwendung (Thompson & Weiss, 2011). Im Rahmen einer solchen Kalibrierungsstudie werden mit einer hinreichend großen Stichprobe die Itemparameter unter Nutzung eines IRT-Modells geschätzt und die Modellpassung der Items geprüft. Items werden auf Basis dieser und weiterer Analysen für den Pool selektiert oder eliminiert (vgl. dazu Kapitel 2, Konstruktion IRT-basierter Tests). Da die Itemparameter im Verlauf der CAT-Anwendung üblicherweise nicht neu geschätzt werden, hat die Kalibrierungsstudie eine besondere Wichtigkeit. Je größer der so zusammengestellte Itempool ist, desto genauer kann die Auswahl und Vorgabe von Items im Verlauf eines CAT an das Antwortverhalten der Testperson j angepasst werden. Ein für CAT optimaler Itempool ist so beschaffen, dass zu jedem Zeitpunkt im Testverlauf, gegeben der aktuellen vorläufigen individuellen Merkmalsschätzung $\hat{\theta}_j$, mindestens ein Item mit hohem diagnostischen Informationsgehalt vorliegt (Reckase, 2009). Daraus folgt, dass ein für CAT optimaler Itempool über verschiedene Bereiche von Merkmalsausprägungen hinweg möglichst viele informative Items enthalten sollte. Liegt ein Rasch-Modell zugrunde, ist ein Item dann besonders informativ, wenn die Itemschwierigkeit b der geschätzten individuellen Merkmalsausprägung $\hat{\theta}_j$ möglichst nahe kommt. Da optimale Itempools eine hinreichende Anzahl von Items erfordern, die breit über das gesamte Spektrum des zu messenden Merkmals verteilt sind, können diese zuweilen Itemzahlen im vierstelligen Bereich verlangen (Frey, 2012). Um brauchbare Itempools mit einem akzeptablen Aufwand konstruieren und kalibrieren zu können, empfiehlt es sich, Monte-Carlo-Simulationsstudien (van der Linden & Glas, 2010) durchzuführen bevor Items entwickelt oder Daten erhoben werden. Solche

Simulationsstudien erlauben es beispielsweise, die Größe und Struktur eines möglichen CAT-Itempools oder die CAT-Länge zu variieren und Effekte auf die Messpräzision zu untersuchen (Thompson & Weiss, 2011). Damit lässt sich eine Balance zwischen der Größe eines angestrebten Itempools und der Messeffizienz finden und somit eine hinreichende, aber nicht übergroße Menge an Items entwickeln. Darüber hinaus empfiehlt es sich im Anschluss an die Itementwicklung und Itempool-Kalibrierung, Simulationsstudien unter Nutzung der in der Kalibrierungsstudie gewonnenen Informationen zu wiederholen. Mithilfe der Ergebnisse kann der angestrebte Test spezifiziert werden oder aber gegebenenfalls entschieden werden, ob weitere Items entwickelt werden sollten, um den Itempool hinsichtlich seiner Struktur und Größe gezielt zu erweitern (Thompson & Weiss, 2011).

3.1.2 Personenparameterschätzung

Im Zuge von linearen, nicht-adaptiven Tests werden Personenparameterschätzungen üblicherweise einmalig nach Testende vorgenommen und basieren auf allen gegebenen Antworten der Testpersonen. Beim adaptiven Testen erfolgt hingegen die Schätzung von Personenparametern auch im Verlauf der Testung. Dabei wird jeweils die Antwort einer Testperson auf ein Item bewertet und die individuelle Merkmalsschätzung aktualisiert. Die vorläufigen, während der Testung geschätzten Personenparameter bilden dann die Basis für die nachfolgende adaptive Itemauswahl. Am Ende des CAT wird die individuelle Merkmalsausprägung der Testperson final geschätzt. Dafür können die üblichen Schätzmethoden im Zusammenhang mit IRT genutzt werden (van der Linden, 2016). Wie in Kapitel 2 beschrieben, lassen sich dabei zwei Ansätze zur Schätzung von Personenparametern unterschieden: Maximum-Likelihood-Schätzer (bspw. MLE; Lord, 1980 oder WLE; Warm, 1989) und Bayesianische Schätzer (bspw. EAP; Bock & Aitkin, 1981; Bock & Mislevy, 1982 oder MAP; Bock & Aitkin, 1981). Bei Bayesianischen Schätzern werden nicht nur die Antworten der zu testenden Person verwendet, sondern

auch Informationen zur Verteilung des interessierenden Merkmals in der Zielpopulation werden in die Personenparameterschätzung einbezogen. Dies führt dazu, dass Bayesianische Schätzer zwar höhere Messpräzision aufweisen, allerdings sind vor allem in den extremen Bereichen der Merkmalsverteilung die Schätzer stärker verzerrt (Frey, 2012). Besonders zur Geltung kommen diese Eigenschaften Bayesianischer Schätzer aber eher bei Tests mit geringer Itemzahl beziehungsweise Testlänge. Werden 20 oder mehr Items vorgegeben, fallen die Unterschiede gering aus (van der Linden, 1998a). Bei invariantem Antwortverhalten (bspw. wenn im Testverlauf alle bisher gegebenen Antworten einer Person richtig oder falsch waren) können nicht alle Schätzmethoden Personenparameter ausgeben. Dies trifft für den MLE zu, während der WLE sowie alle Bayesianischen Ansätze einen Schätzwert trotz invariantem Antwortmuster liefern können.

Obwohl sich, wie beschrieben, Unterschiede bezüglich der Schätzgüte zwischen Maximum-Likelihood- und Bayesianischen Schätzern zeigen, können beide Methoden zur Schätzung individueller Merkmalsausprägungen herangezogen werden (zu Vor- und Nachteilen unterschiedlicher Personenparameterschätzer siehe bspw. Van der Linden & Pashley, 2010 oder P. E. Cheng & Liou, 2000).

3.1.3 Itemauswahl

Im Zentrum adaptiver Tests steht die Itemauswahl zu Beginn und im Verlauf der Testung (Frey & Seitz, 2009). Dabei gilt es einerseits zu bestimmen, wie das erste Item im Test ausgewählt wird, also zu einem Zeitpunkt an dem die Testperson noch kein Antwortverhalten gezeigt hat, und andererseits, welches Kriterium für die nachfolgend auszuwählenden Items verwendet werden soll (Frey, 2012). Im adaptiven Testverlauf wird per Definition die Auswahl des nachfolgend vorzulegenden Items auf Basis des zuvor gezeigten Antwortverhaltens einer Testperson vorgenommen. Beim Start des CAT wurde der Testperson allerdings noch kein Item vorgelegt und somit noch keine Antwort abgegeben, die zur Schätzung des individuellen Personenparameters und zur Itemauswahl

genutzt werden könnte. Liegen zu Testbeginn also keine Informationen zur individuellen Merkmalsausprägung der Testperson vor, wird häufig ein Item zufällig oder ein Item mit mittlerer Schwierigkeit als Startitem vorgegeben. Manchmal werden auch sehr leichte, sogenannte Eisbrecheritems, zu Beginn der Testung administriert, die einen problemlosen Einstieg in den Test gewährleisten sollen (Frey, 2012). Sind Informationen im Hinblick auf die zu messende individuelle Merkmalsausprägung einer Testperson verfügbar, beispielsweise aus früheren Testungen, aus Testungen mit vergleichbaren Instrumenten oder über Korrelationen mit anderen bekannten Merkmalsausprägungen, so können diese für die Auswahl des ersten zu administrierenden Items im CAT genutzt werden. Auf Basis solcher Vorinformationen kann bereits eine vorläufige Schätzung des individuellen Personenparameters vorgenommen und ein Item mit optimalen Eigenschaften aus dem Itempool ausgewählt und vorgegeben werden (van der Linden, 1999). Vor allem bei kürzen Testlängen kann die Auswahl des Startitems einen Einfluss auf die Messpräzision haben, der aber bei zunehmender Testlänge deutlich abnimmt (Hambleton, Zaal & Pieters, 1991).

Bei der Itemauswahl im CAT-Verlauf können die auf Basis der bis dato administrierten Items gewonnenen Informationen genutzt werden, um das Item aus der Menge aller noch zur Verfügung stehender Items auszuwählen, welches besonders viel diagnostische Information über die individuelle Merkmalsausprägung der Testperson liefert. Hierfür können verschiedene Methoden gewählt werden. Allen gemein ist, dass unter Berücksichtigung der aktuellen Personenparameterschätzung ein zuvor bestimmtes statistisches Kriterium maximiniert oder minimiert wird. Im Hinblick auf diese *Optimalitätskriterien* lassen sich zwei Ansätze unterscheiden: die Itemauswahl nach Iteminformation und die Bayesianische Itemauswahl. Bei der Bayesianischen Itemauswahl wird die A-posteriori-Verteilung der Personenparameter berücksichtigt. Einen Überblick

verschiedener Ansätze bieten beispielsweise van der Linden (1998a), Veldkamp (2010) oder van der Linden und Pashley (2010).

Das am stärksten verbreitete und am häufigsten angewandte Kriterium zur Itemauswahl im CAT-Verlauf ist das der maximalen Information, bei dem – gegeben der aktuellen individuellen Merkmalsschätzung $\hat{\theta}_j$ – das Kandidatenitem i^* ausgewählt wird, welches den höchsten Wert an Information $I_{i^*}\left(\theta_j\right)$ aufweist, also über maximale Information verfügt (Lord, 1980). Praktisch wird für alle bisher nicht präsentierten Items deren Informationsfunktion am Punkt $\hat{\theta}_j$ berechnet und das Item mit dem höchsten Wert zur Administration ausgewählt. Dabei ist die Berechnung der Iteminformationsfunktion (IIF, siehe Kapitel 2) je nach genutztem IRT-Modell unterschiedlich. Wird ein 1PL-Modell verwendet, ergibt sich die IIF eines Kandidatenitems i^* entsprechend Formel 2.5 aus:

$$I_{i^*}\left(\theta_j\right) = p_{i^*}(\theta) q_{i^*}(\theta), \tag{3.1}$$

wobei $p_{i^*}(\theta)$ die Wahrscheinlichkeit ausdrückt, mit der eine zu testenden Person mit der latenten Merkmalsausprägung θ das zur Administration verfügbare Item i^* korrekt löst. Mit $q_{i^*}(\theta)$ wird die entsprechende Gegenwahrscheinlichkeit ausgedrückt. So wird jeweils das Item ausgewählt, dessen Itemschwierigkeit die geringste Abweichung zum vorläufig geschätzten Personenparamter aufzeigt. Bei Verwendung von 2PL- oder 3PL-Modellen fließen in die Berechnung der Iteminformation neben der Itemschwierigkeit auch Diskriminationsparameter und/oder Pseudo-Rateparameter direkt mit ein. Dies kann dazu führen, dass nicht das Item ausgewählt wird, welches am Nächsten zur vorläufig geschätzten Merkmalsausprägung ist. Trotzdem werden auch bei CAT auf Basis von 2PL- oder 3PL-Modellen meist Items ausgewählt, die für die Testperson eine ungefähr mittlere Schwierigkeit aufweisen (Frey, 2012).

Im Verlauf der Testung werden die Veränderungen in den vorläufigen Personenparameterschätzungen immer geringer und damit die geschätzten Werte stabiler. Zudem verringert sich im Testverlauf der Standardfehler der Personenparameterschätzung. Bei der Itemauswahl gemäß des Optimalitätskriteriums der maximalen Information verringert sich der Standardfehler der Personenparameterschätzung mit jeder Itemauswahl in größtmöglicher Weise (aufgrund des in Kapitel 2, Formel 2.7 dargestellten Zusammenhangs von Standardfehler und Iteminformation). Im Vergleich mit anderen Optimalitätskriterien führt die Nutzung des Kriteriums der maximalen Iteminformation meist zu vergleichbar guten Ergebnissen und schneidet nur bei besonderen Anwendungssituationen, wie beispielsweise sehr kurzen Tests, geringfügig schlechter ab (Veerkamp & Berger, 1997; Chang & Ying, 1996).

3.1.4 Nebenbedingungen an die Testzusammenstellung

Durch die zuvor beschriebene Itemauswahl auf Basis eines rein statistischen Optimalitätskriteriums bleiben mögliche nicht-statistische Nebenbedingungen an die Zusammenstellung eines CAT unberücksichtigt. Der Umgang mit Nebenbedingungen wird als *Constraint-Management* bezeichnet und kann sich auf Methoden zur Regelung inhaltlicher Anforderungen an die Testzusammenstellung (engl. *Content-Management*) und auf Methoden zur Kontrolle der Vorgabehäufigkeit von Items (engl. *Item-Exposure-Control*) beziehen. Die Festsetzung von Nebenbedingungen an die Testzusammenstellung im CAT ist allerdings ein optionales Element, kommt also nicht zwangsläufig bei allen CAT zur Anwendung. Wenn bei der praktischen Anwendung von CAT die Itemauswahl durch Nebenbedingungen eingeschränkt wird, gilt es einen Kompromiss zwischen dem in Abschnitt 3.1.3 beschriebenem statistischen Optimum und der Erfüllung nicht-statistischer Anforderungen zu finden (Frey, 2012). Um gesetzten Nebenbedingungen gerecht zu werden, kann es vorkommen, dass nicht immer Items mit dem höchsten diagnostischen Informationsgehalt gewählt werden. So kann die Berücksichtigung nicht-

statistischer Anforderung bei der Itemauswahl Einbußen in der Messpräzision nach sich ziehen.

Content-Management. Soll beispielsweise im Verlauf eines CAT jeder Testperson ein bestimmter Anteil von Items eines Inhaltsbereichs, einer Subfacette, eines spezifischen Antwortmodus oder einer bestimmten Darstellungsform vorgegeben werden, gilt es inhaltliche Nebenbedingungen an die Testzusammenstellung auf Individualebene mittels Content-Management-Methoden zu berücksichtigen. Oftmals werden mehrere solcher Nebenbedingungen formuliert, die simultan Berücksichtigung finden müssen, wobei ein Item mehrere dieser Nebenbedingungen erfüllen kann. Heuristische Verfahren setzen üblicherweise eine Form der Gewichtung der Informationsfunktion von Items um, die den Nebenbedingungen am ehesten gerecht werden. Bekannte heuristische Verfahren sind beispielsweise das Weighted-Deviation-Model (WDM; Stocking & Swanson, 1993), das Weighted-Penalty-Model (WPM; Shin, Chien, Way & Swanson, 2009) sowie der Maximum-Priority-Index (MPI; Y. Cheng & Chang, 2009). Der MPI wird im Rahmen dieser Arbeit eingesetzt (Kapitel 7) und soll daher nachfolgend kurz eingeführt werden.

Der MPI basiert auf einer Constraint-Matrix $\mathbf{C}$, deren Größe sich aus der Anzahl der Items im Itempool multipliziert mit der Anzahl zu erfüllender Nebenbedingungen ergibt. Mit $c_{i^*k} = 1$ wird indiziert, dass ein verfügbares Item i^* zur Erfüllung einer Nebenbedingung k relevant ist, während $c_{i^*k} = 0$ ausdrückt, dass dies nicht der Fall ist. Der Priority-Index PI_{i^*} eines für die Administration zur Verfügung stehenden Items i^* kann folgendermaßen berechnet werden:

$$\mathrm{PI}_{i^*} = I_{i^*} \prod_{k=1}^{K} \left(w_k f_k \right)^{c_{i^*k}}, \tag{3.2}$$

wobei I_{i^*} die Iteminformationsfunktion für das Item gegeben der aktuellen Schätzung der individuellen Merkmalsausprägung $\hat{\theta}$ ausdrückt. Zudem steht w_k für das Gewicht der

Nebenbedingung k, das genutzt wird, um die Relevanz verschiedener Nebenbedingungen abzubilden. Schließlich wird durch f_k ausgedrückt, wie dringlich die Erfüllung der Nebenbedingung zum aktuellen Zeitpunkt im CAT-Verlauf ist. Die Berechnung von f_k ergibt sich durch:

$$f_k = \frac{\left(X_k - x_k\right)}{X_k}, \tag{3.3}$$

dabei entspricht X_k der Anzahl von Items, die benötigt werden, um die Nebenbedingung k zu erfüllen, und x_k die Anzahl solcher schon administrierter Items. Sollte zum aktuellen Testzeitpunkt noch kein Item administriert worden sein, das die Nebenbedingung k erfüllt, gilt $\left(x_k = 0\right)$ und der Term nimmt den Wert 1 an. Daraus folgt, dass die Iteminformationsfunktion eines Items, welches diese Nebenbedingung erfüllt, stark gewichtet wird. Der Wert von f_k nimmt ab, je mehr Items administriert wurden, die die Nebenbedingung k erfüllen, bis genügend entsprechende Items im CAT vorgegeben wurden $\left(x_k = X_k\right)$ und $f_k = 0$ wird. Ist ein Item i^* nicht geeignet, um die Nebenbedingung k zu erfüllen, ist $c_{i^*k} = 0$ und der Term $w_k f_k$ hat keinen Einfluss auf den PI_{i^*}. Im Verlauf der Itemauswahl unter Nutzung des MPI zum Umgang mit Nebenbedingungen an die Testzusammenstellung wird zunächst für jedes noch verfügbare Item im Itempool der PI berechnet und anschließend das Item mit dem höchsten PI zur Administration ausgewählt.

Ein anderes, nicht-heuristisches Verfahren, welches aktuell häufig angewandt wird und als besonders leistungsstark gilt, ist der Shadow-Test-Approach (STA; van der Linden & Reese, 1998; van der Linden, 2005). Dabei wird die Itemauswahl unter Nebenbedingungen als Problem der eingeschränkten kombinatorischen Optimierung verstanden und gelöst. Der Kernpunkt, der den STA von anderen Content-Management-Methoden unterscheidet, ist, dass Items nicht direkt aus dem Itempool ausgewählt werden,

sondern aus einem zuvor zusammengestellten Shadow-Test. Anders als bei heuristischen Methoden wird beim STA die Iteminformation in Bezug auf die aktuelle Personenparameterschätzung als Zielfunktion für die Optimierung unter Berücksichtigung der gesetzten Nebenbedingungen an die Testzusammenstellung angesehen (van der Linden, 2005). Für einen mit dem STA administrierten CAT ergeben sich die folgenden (verkürzt dargestellten) Schritte: (1) Schätzung des Personenparameters; (2) Zusammenstellung eines Shadow-Tests, der alle Nebenbedingungen erfüllt und die Zielfunktion optimiert; (3) Auswahl eines Items aus dem Shadow-Test, das gegeben der aktuellen Personenparameterschätzung die meiste diagnostische Information liefert; (4) Aktualisierung der Personenparameterschätzung; (5) Wiederholung der Schritte 2 bis 4 bis das definierte Abbruchkriterium erfüllt ist (He, Diao & Hauser, 2014). Dabei beinhalten Shadow-Tests, die im CAT Verlauf erstellt werden, immer auch alle der Person schon vorgelegten Items, die dann als zusätzliche Einschränkungen in der Testzusammenstellung verstanden werden können (Choi, Moellering, Li & van der Linden, 2016). Es wird deutlich, dass die wiederholte Lösung dieses Optimierungsproblems rechenintensiv ist. Zudem ist lineare Programmierung und spezielle Solver-Software erforderlich. Allerdings liefert der STA auch bei einer hohen Anzahl von gesetzten Nebenbedingungen in der Regel gute Lösungen (Frey, 2012; für einen Vergleich verschiedener heuristischer Methoden und des STA siehe He et al., 2014).

Item-Exposure-Control. Bei der Itemauswahl durch Verwendung eines statistischen Optimalitätskriteriums kann es dazu kommen, dass einzelne oder mehrere Items sehr vielen bis allen Testpersonen (engl. *Overexposure*) oder nur sehr wenigen bis keinen Testpersonen (engl. *Underexposure*) zur Bearbeitung vorgelegt werden. Overexposure einzelner Items ist meist nicht wünschenswert, da auf diese Weise zum Beispiel Iteminhalte schneller bekannt und diese Items dann oft unbrauchbar werden. Dies ergibt sich daraus, dass die Itemschwierigkeit abnehmen kann (bspw. weil Testpersonen

das Item und die Lösung gelernt haben) und es fraglich ist, ob durch das Item noch die intendierten Inhalte gemessen werden (oder stattdessen bspw. Gedächtnisleistung). Underexposure von Items gilt es ebenso möglichst zu vermeiden, da die Itementwicklung meist ein zeit- und kostenintensiver Prozess ist und der erarbeitete Itempool möglichst umfänglich ausgenutzt werden sollte (Gu & Reckase, 2007). Strategien zur Kontrolle der Itemvorgabehäufigkeit werden als Exposure-Control bezeichnet (Frey, 2012). Für CAT gehören hierzu beispielsweise Randomisierungsansätze (bspw. Kingsbury & Zara, 1989) oder die Sympson-Hetter-Methode (Sympson & Hetter, 1985). Einen Überblick zu verschiedenen Exposure-Control-Strategien geben beispielsweise Georgiadou, Triantafillou und Economides (2007) oder Leroux, Lopez, Hembry und Dodd (2013).

3.1.5 Abbruchkriterium

Wie in Abbildung 3.1 dargestellt werden im Verlauf eines CAT die Schritte (1) Itemauswahl, (2) Antwortregistrierung und -bewertung und (3) Personenparameterschätzung solange wiederholt, bis ein oder eine Kombination aus mehreren zuvor definierten Abbruchkriterien erfüllt ist. Besonders häufig genutzte Abbruchkriterien legen die Beendigung eines CAT nach Vorgabe einer festgelegten Anzahl von Items, nach Erreichen einer bestimmten Messpräzision oder nach Ablauf einer zuvor definierten maximalen Testzeit fest (Wainer, 2000). Zudem kann ein CAT nur fortgesetzt werden, solange noch Items im Itempool zur Verfügung stehen (Linacre, 2000). Darüber hinaus gibt es weitere mögliche Abbruchkriterien, beispielsweise solche, die sich auf die erwartete Veränderung der Personenparameterschätzung (Babcock & Weiss, 2009) oder des Standardfehlers der Parameterschätzung beziehen (Choi, Grady & Dodd, 2011). Dies liegt darin begründet, dass eine Weiterführung des CAT ohne substantielle Veränderung der Schätzwerte beziehungsweise Verbesserung der Schätzgenauigkeit die Testzeit nur unnötig erhöhen würde. Die Wahl eines geeigneten Abbruchkriteriums sollte von dem jeweiligen Anwendungskontext, der Größe und

Struktur des zur Verfügung stehenden Itempools und eventuellen Nebenbedingungen an die Testzusammenstellung bestimmt werden. Sollten die aus dem CAT resultierenden Schätzungen der individuellen Merkmalsausprägungen von Testpersonen für individualdiagnostische Zwecke (bspw. für interindividuelle Vergleiche) genutzt werden, bietet es sich an, als Kriterium das Erreichen einer bestimmten Messpräzision zu nutzen, da vergleichbar präzise Personenparameterschätzungen erstrebenswert sind (Frey, 2012). Hingegen sind CAT mit flexibler Testlänge bei der Untersuchung von Gruppen und der gleichzeitigen Testung vieler Personen, wie beispielsweise im Rahmen von LSAs, oftmals nicht praktikabel. Daher werden in diesem Kontext eher Abbruchkriterien nach fester Testlänge oder Testzeit genutzt. Dadurch entstehende Unterschiede in der Präzision der individuellen Personenparameterschätzungen können bei einer Auswertung auf aggregierter Ebene als weniger problematisch angesehen werden. Oftmals wird allerdings auch eine Kombination aus mehreren Abbruchkriterien genutzt (Frey, in Druck). Ein Vergleich verschiedener Abbruchkriterien findet sich beispielsweise bei Babcock und Weiss (2009) oder C. Wang, Weiss und Shang (2019).

3.2 Multidimensionales adaptives Testen

Der zuvor beschriebene Ablauf eines CAT sowie seine bestimmenden Element bezogen sich vor allem auf den Einsatz eindimensionaler Tests. Wird allerdings angenommen, dass das gezeigte Antwortverhalten einer Testperson nicht nur auf die Ausprägung eines latenten Merkmals zurückzuführen ist, sondern vielmehr mehrere latente Merkmale oder Merkmalsdimensionen ursächlich für das beobachtete Antwortverhalten sind, dann sollte multidimensionales adaptive Testen (MAT; Segall 1996; Frey & Seitz, 2009) eingesetzt werden. Damit können individuelle Ausprägungen mehrerer Merkmale oder Merkmalsdimensionen simultan gemessen werden. Dabei entspricht der Grundgedanke des multidimensionalen adaptiven Testens dem des

eindimensionalen CAT. Allerdings liegt der Vorzug von MAT in dem hohen Maß an diagnostischer Information, die bei der simultanen Messung korrelierter Merkmale oder Merkmalsdimensionen erschlossen werden kann. Zudem bietet MAT eine erhöhte Messeffizienz im Vergleich zu linearen, nicht-adaptiven Tests oder mehreren eindimensionalen CAT (Frey & Seitz, 2010).

Als psychometrische Modelle werden beim MAT mehrdimensionale IRT-Modelle (MIRT-Modelle, siehe Kapitel 2) genutzt. Die Personenparameterschätzung im MAT kann sowohl durch die Verwendung von Maximum-Likelihood-Schätzern als auch durch Bayesianischen Schätzern erfolgen. Wie in Abschnitt 3.1.2 beschrieben, liegen die Vorzüge der Bayesianischen Schätzung in der Nutzung der, im mehrdimensionalen Fall, multivariaten Merkmalsverteilung sowie der Schätzbarkeit von Personenparametern trotz invariatem Antwortmuster. Eine Darstellung von Vor- und Nachteilen unterschiedlicher Personenparameterschätzer für MAT findet sich bei Diao und Reckase (2009).

Die multidimensionale Struktur eines zu messenden Merkmals oder auch die korrelative Beziehung verschiedener simultan zu messender Merkmale kann im Rahmen von adaptiven Tests beziehungsweise in adaptiven Algorithmen auf verschiedene Art berücksichtigt werden: bei der Itemauswahl (zu Testbeginn und im Testverlauf) und bei der Schätzung von Personenparametern (im Testverlauf und zum Testende). In Studie I und II der vorliegenden Arbeit (Kapitel 6 und 7) werden adaptive Algorithmen, die Multidimensionalität in verschiedener Weise berücksichtigen, erprobt und verglichen. Ein häufig genutzter MAT-Ansatz, der im Rahmen dieser Arbeit auch Anwendung findet, verbindet Bayesianische Personenparameterschätzungen und Itemauswahl nach dem von Segall (1996) vorgeschlagenen Bayesianischen Optimalitätskriterium (bspw. Frey & Seitz, 2010; W.-C. Wang & Chen, 2004). In den folgenden beiden Abschnitten soll daher kurz auf die für diese Arbeit relevanten Aspekte der Itemauswahl und des Constraint-Managements in MAT eingegangen werden.

3.2.1 Itemauswahl

Wie beim eindimensionalen adaptiven Testen basieren auch bei MAT Itemauswahlmethoden auf der Maximierung oder Minimierung eines statistischen Kriteriums zum Zeitpunkt der aktuellen Schätzung des $\boldsymbol{\theta}$-Vektors (Reckase, 2009). Verschiedene Optimalitätskriterien eindimensionaler adaptiver Tests wurden für den multidimensionalen Fall erweitert (für Beispiele siehe C. Wang & Chang, 2011). Grundsätzlich wird mit der Itemauswahl immer eine Reduktion der erwarteten Varianz bei der Schätzung des vorläufigen multidimensionalen Merkmalsvektors $\hat{\boldsymbol{\theta}}$ verfolgt (Frey & Seitz, 2009). Ein häufig angewendetes Optimalitätskriterium für MAT ist der von Segall (1996) vorgestellte Bayesianische Ansatz. Bei diesem wird berücksichtigt, dass die Antwort einer Person auf ein Item nicht nur Information hinsichtlich der Merkmalsdimension liefert, die das betreffende Item abbildet, sondern auch hinsichtlich anderer, korrelativ verbundener Merkmalsdimensionen. Für die Itemauswahl wird daher die A-posteriori-Verteilung der Merkmalsausprägung $\boldsymbol{\Phi}$ genutzt. Im adaptiven Test wird das Kandidatenitem i^* aus dem Itempool der (noch) zur Verfügung stehenden Items ausgewählt, welches den folgenden Ausdruck maximiert:

$$\left|\mathbf{I}\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_T\right)+\mathbf{I}\left(\boldsymbol{\theta},u_{i*}\right)+\boldsymbol{\Phi}^{-1}\right|. \tag{3.4}$$

In das zu maximierende Kriterium fließt so die Informationsmatrix der T bereits administrierten Items $\mathbf{I}\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_T\right)$, die Informationsmatrix des zur Administration verfügbaren Items i^* $\mathbf{I}\left(\boldsymbol{\theta},u_{i*}\right)$ und die Inverse der A-posteriori-Verteilung der Merkmalsausprägung $\boldsymbol{\Phi}^{-1}$ ein. Dadurch wird dasjenige Item zur Vorgabe ausgewählt, welches die größte Messpräzisionssteigerung hinsichtlich aller zu messenden Dimensionen bietet. Bei Verwendung des vorgestellten Bayesianischen Ansatzes konnten W.-C. Wang und Chen (2004) zeigen, dass MAT im Vergleich zu linearen, nicht-adaptiven und auch zu

mehreren eindimensionalen CAT effizienter ist, je mehr Dimensionen simultan gemessen werden und je stärker diese korrelieren. Diese Messeffizienzsteigerung fällt andererseits aber geringer aus, wenn der zur Verfügung stehende Itempool suboptimal ist und zu viele Restriktionen vorgegeben sind (Frey, 2012).

3.2.2 Nebenbedingungen an die Testzusammenstellung

Zur Berücksichtigung von Nebenbedingungen für die Testzusammenstellung beim MAT sind Erweiterungen von Constraint-Management-Methoden zum Content-Management und zur Exposure-Control erforderlich. Beispielsweise wurde die Sympson-Hetter-Methode (Sympson & Hetter, 1985) als Exposure-Control-Strategie für den multidimensionalen Fall erweitert (Yao, 2014). Der STA (van der Linden & Reese, 1998) wurde von Veldkamp und van der Linden (2002) ebenfalls für die Administration von MAT erweitert. Zum Content-Management wurde zudem das WPM (Shin et al., 2009) zum Multidimensional-Weighted-Penalty-Model (MPWM; Born & Frey, 2017) erweitert und der zuvor näher vorgestellte MPI (siehe Abschnitt 3.1.4, Formel 3.2) zum Multidimensional-Maximum-Priority-Index (MMPI; Frey, Cheng & Seitz, 2011; Frey, Seitz & Kröhne, 2013) weiterentwickelt. Dabei wurde die für eindimensionale CAT genutzte Iteminformationsfunktion durch das von Segall (1996) eingeführte Bayesianische Informationskriterium (Formel 3.4) ersetzt. Trotz dieser Änderung bleibt das zugrundeliegende Prinzip des MMPI gleich dem des MPI. Der Priority-Index PI_{i^*} eines zur Administration verfügbaren Items i^* für den multidimensionalen Fall ergibt sich folgendermaßen:

$$\mathrm{PI}_{i^*} = \left| \mathbf{I}\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_T\right) + \mathbf{I}\left(\boldsymbol{\theta}, u_{i^*}\right) + \boldsymbol{\Phi}^{-1} \right| \prod_{k=1}^{K} \left(w_k f_k\right)^{c_{i^*k}} . \tag{3.5}$$

Wobei (wie bereits im Zuge der Ausführungen zu Formel 3.2 dargelegt wurde) w_k die Relevanz einer Nebenbedingung und f_k die Dringlichkeit der Erfüllung der Nebenbedingung zum aktuellen Zeitpunkt im MAT-Verlauf abbildet.

In diesem Kapitel wurden die Grundidee sowie der Ablauf eines CAT dargestellt, um die Basis der in den Studien I und II (Kapitel 6 und 7) dieser Arbeit angewendeten multidimensionalen adaptiven Algorithmen zu legen. Mit Hilfe der Studien sollten Hinweise generiert werden, wie das komplexe Konstrukt der ICT-Skills (dargestellt in Kapitel 4) mit einem adaptiven Test im Rahmen von LSAs erfasst werden könnte (Fragestellungen der Studien werden in Kapitel 5 hergeleitet).

4 Informations- und Kommunikationstechnologie-bezogene Fertigkeiten und Fähigkeiten

In diesem Kapitel wird nach den zuvor beschriebenen methodisch statistisch relevanten Grundlagen der vorliegenden Arbeit die inhaltliche Basis der später dargestellten Studien erläutert. Dazu wird zunächst allgemein auf die Bedeutung von Fertigkeiten und Fähigkeiten im Umgang mit Informations- und Kommunikationstechnologien sowie Ansätze und Herausforderungen zur Messung dieser (Kapitel 4.1) und anschließend spezifisch auf das Projekt CavE-ICT (Kapitel 4.2 und 4.3) eingegangen.

4.1 Bedeutung von ICT-Skills und deren Messung

Digitale Technologien sind in unserer globalisierten Informationsgesellschaft mit dem täglichen Leben fest verbunden und nicht mehr wegzudenken (Zabal et al., 2013). Die Schnelligkeit technologischer Entwicklungen und das Fortschreiten der Technisierung in allen Lebens- und Arbeitsbereichen führen dazu, dass die Bedeutung von Fertigkeiten und Fähigkeiten im Umgang mit Informations- und Kommunikationstechnologien (engl. *Information and Communication Technologies*, ICT) auch stetig zunimmt. Fertigkeiten und Fähigkeiten im Umgang mit ICT, das heißt für das Lösen von Aufgaben in ICT-Umgebungen, werden im Folgenden als *ICT-Skills* bezeichnet (Engelhardt, Goldhammer, Naumann & Frey, 2017). ICT-Skills können zudem als Kernkompetenz zur Gewährleistung lebenslangen Lernens gesehen werden (bspw. Ferrari, Punie & Redecker, 2012; Poynton, 2005), da viele gesellschaftlich relevante Wissensbereiche über die Lebensspanne weitgehend selbstgesteuert über digitale Medien erworben werden können (Ezziane, 2007). ICT-Skills stellen mittlerweile ein wichtiges Bildungsziel dar, denn der kompetente Umgang mit Computertechnologien ist sowohl im privaten, im

bildungsbezogenen wie auch im beruflichen Kontext von hoher Relevanz für eine erfolgreiche Teilhabe an modernen Wissensgesellschaften (bspw. van Deursen & van Dijk, 2011; Bundesministerium für Bildung und Forschung, 2016; Binkley et al., 2012; Eshet-Alkalai, 2004; International ICT Literacy Panel, 2002; Fraillon et al., 2013). Somit können ICT-Skills auch als „Kulturtechnik" (Bos et al., 2014) verstanden werden, die eingesetzt wird um Ziele persönlicher, beruflicher und sozialer Natur zu erreichen (Kultusminister Konferenz, 2016). Damit wird impliziert, dass ein geringes Maß an ICT-Skills dazu führen kann, dass Menschen Benachteiligungen und möglicherweise nicht im vollen Umfang gesellschaftliche Teilhabe erfahren. Um solchen Benachteiligungen entgegenzuwirken soll bereits im Bereich der schulischen Bildung der Umgang mit Computern als Informations- und Kommunikationsmedium oftmals als transversale Fertigkeit in traditionelle Unterrichtsfächer integriert werden (Kultusminister Konferenz, 2016). Inwiefern es tatsächlich gelingt den kompetenten Umgang mit ICT im Schulkontext zu vermitteln und sozial bedingten Disparitäten im Kompetenzerwerb (Livingstone & Helsper, 2007) kompensierend entgegenzuwirken, ist noch weitestgehend unklar. Genau solchen Fragestellungen, der sozialen Ungleichheit im Bildungssystem gehen häufig LSAs, als wesentliche Instrumente des Bildungsmonitorings nach (bspw. Gniewosz & Gräsel, 2015). Bei LSAs handelt es sich um groß angelegte (engl. Large Scale) nationale und internationale Untersuchungen die die Beurteilung von Personen (engl. Assessment) hinsichtlich bestimmter Merkmale zum Ziel haben. Von Interesse ist dabei die empirische Erfassung des Gesamtbilds einer Population (ggf. im Vergleich mit anderen Populationen) und nicht die Beschreibung des Einzelnen im Sinne einer Individualdiagnostik (Kubinger, 2014).

Es liegt somit nahe, dass die Erfassung von ICT-bezogenen Konstrukten auch und besonders im Rahmen von LSAs fokussiert wird. Im Rahmen des *Nationalen Bildungspanels* (German National Educational Panel Study – NEPS; Blossfeld & Roßbach, 2019) wird beispielsweise Engagement in digitalen Lernumwelten erfragt (Kuger, Linberg, Bäumer &

Struck, 2018), was als wesentliche Voraussetzung für die (Weiter-)Entwicklung von ICT-Skills verstanden werden kann (Goldhammer, Gniewosz & Zylka, 2016). In NEPS werden ebenfalls ICT-Skills erfasst, seit 2018 auch durch computerbasierte und interaktive Items, die die in der ersten Projektphase bestehende Erfassung über Multiple-Choice-Items, die möglichst realitätsnah und mit Screenshots angereichert dargeboten wurden, ergänzen (Leibniz-Institut für Bildungsverläufe e.V., 2018). Seit dem Jahr 2000 werden in der Erhebung des *Programme for International Student Assessment* (PISA; OECD, 2014) Variablen zu Verfügbarkeit und Nutzung sowie Motivation in Bezug auf ICT als optionaler Fragebogen erhoben (OECD, 2013). Beispielsweise wird im Rahmen des *Programme for the International Assessment of Adult Competencies* (PIAAC; Rammstedt, 2013) technologiebasiertes Problemlösen von Erwachsenen untersucht. Zudem werden im Rahmen der *International Computer and Information Litracy Study* (ICILS; Eickelmann et al., 2019) Schülerinnen und Schüler der achten Jahrgangsstufe hinsichtlich ihrer computer- und informationsbezogenen Fertigkeiten untersucht. Dabei wird neben ICT-Skills über den Einsatz von Multiple-Choice- und simulationsbasierten Items auch die Motivation ICT zu nutzen erhoben (Senkbeil, 2018).

Da die künftige Testrealität bei LSAs in dynamischen, interaktiven und adaptiven, computerisiert administrierten Erhebungen liegt, sind vor allem verhaltensnahe Messungen von ICT-Skills erst ermöglicht. Dennoch bleibt zu konstatieren, dass noch immer eher distale Maße von ICT-Skills, wie subjektive Einschätzungen oder reines Wissen, im Rahmen von LSAs aber auch darüber hinaus, erhoben werden (Goldhammer et al., 2016; Kuger et al., 2018; Richter, Naumann & Horz, 2010; Markauskaite, 2007). Dabei kann einerseits die subjektive Einschätzungen der eigenen Leistung oder Fähigkeit zwar das tatsächliche leistungsthematische Verhalten und damit auch die Leistung bedingen (Eccles, 2006; Ehmke & Siegle, 2008; Heckhausen & Heckhausen, 2010), erlaubt aber keine direkte Übertragung von Selbsteinschätzung zu tatsächlichem

Leistungsvermögen. Es kann also als fraglich gelten, ob aufgrund subjektiver Selbsteinschätzungen valide Aussagen über individuelle Kompetenzunterschiede gemacht werden können. Das Problem bei Wissensfragen andererseits ergibt sich daraus, dass kompetentes Handeln nach Hartig und Klieme (2006) nicht nur Wissen, sondern auch Können erfordert. Reines technikbezogenes Wissen reicht zur kompetenten Bewältigung von Aufgaben in ICT-Umgebungen oftmals nicht aus. Bei der Informationssuche im Internet liegt die Herausforderung beispielsweise nicht nur in der Bedienung einer Suchmaschine, sondern auch und vor allem in der Bewertung der gefundenen Information hinsichtlich ihrer Qualität und Vertrauenswürdigkeit, da im Internet keine Qualitätskontrollmechanismen greifen (Rieh, 2002). So zeigen gerade jüngere Personen, dass die Navigation und Orientierung im Internet keine große Herausforderung für sie darstellt, hingegen zeigen sie Defizite beim Bewerten von im Netz gefunden Informationen (Eshet Alkali & Amichai-Hamburger, 2004; Eshet-Alkalai & Chajut, 2010; Lorenzen, 2001; van Deursen & van Dijk, 2009). Versteht man ICT-Skills demnach als prozedurales Handlungswissen, also die Fähigkeit Wissensbestände in realitätsnahen Situationen anzuwenden, würde ein reiner Wissenstest dem Anspruch, ICT-Skills valide zu messen, nicht gerecht werden (Senkbeil & Ihme, 2019).

Die Bedeutung von theoretischen Konzeptionen zur detaillierten Beschreibung von ICT-Skills sowie die Entwicklung von theoriebasierten Instrumenten zu deren Erfassung haben in den letzten Jahren stetig zugenommen. Die Entwicklung geeigneter computerbasierter und verhaltensnaher Testinstrumente stellt dabei eine besondere Herausforderung in diesem Forschungsbereich dar. Dabei erscheint es am sinnvollsten, ICT-Skills computerisiert zu erheben, da so der tatsächliche Umgang mit ICT in der Testsituation geprüft werden kann. Auf diese Weise wird eine sehr gute Konstruktrepräsentation ermöglicht (Sireci & Zenisky, 2006), was zentral für eine valide Interpretation von Testwerten ist (Goldhammer et al., 2014). Dabei können Items, die den

Umgang mit alltäglich gebräuchlicher und bekannter Software adressieren, zwar sehr authentisch sein, diese Applikationen sind aber nicht einfach in größere Assessments einzubinden (Parshall, Spray, Kalohn & Davey). Ähnliche Applikationen zu simulieren (sozusagen in Items nachzubilden) ist in der Entwicklung recht aufwändig und Itementwickler sind mit der Frage konfrontiert, welche Aspekte in der Simulation abgebildet werden müssen und welche weggelassen werden können (Mislevy, 2013). Je mehr Aspekte durch solche simulationsbasierten Items abgedeckt werden, desto mehr Interaktions- und Explorationsmöglichkeiten bieten sich der Testperson was Itembearbeitungs- und Testzeiten deutlich verlängern kann (Greiff et al., 2013). Der Spagat zwischen einer authentischen und dennoch ökonomisch realisierbaren Itemdarstellung ist für den Implementierungsprozess verhaltensnaher Messungen in computerisierteren Erhebungen von entscheidender Bedeutung (Engelhardt et al., eingereicht).

4.2 Die Cave-ICT-Framework- und Itementwicklung

Das Projekt „Computergestützte, adaptive und verhaltensnahe Erfassung Informations- und Kommunikationstechnologie-bezogener Fertigkeiten (ICT-Skills)“ (CavE-ICT) widmete sich der theoriegeleiteten Entwicklung und empirischen Erprobung eines computerbasierten, verhaltensnahen und interaktiven sowie potentiell adaptiven Tests zur Messung von ICT-Skills. Das so entwickelte Instrument sollte explizit auch im Kontext von LSAs beispielsweise bei künftigen PISA-Erhebungen als internationale Option, Verwendung finden können (Wenzel et al., 2015). Die im Rahmen dieses Projekts verfolgten Ziele, die entwickelten sowie erprobten Items und erhobenen Daten liefern die Grundlage der in dieser Arbeit in insgesamt drei Studien (Kapitel 6, 7 und 8) bearbeiteten Fragestellungen (Kapitel 5), die allerdings deutlich über die konkret formulierten Projektziele hinausgehen. Aus diesem Grund wird im Folgenden das im Projektverlauf

entwickelte Framework zur Beschreibung von ICT-Skills (4.2.1), die darauf aufbauende Itementwicklung (4.2.2) sowie die Datenerhebung im Zuge des CavE-ICT-Feldtests und zentrale Ergebnisse (4.2.3 und 4.2.4) kurz vorgestellt. Weiterführende Informationen zum im Projekt aufgestellten *CavE-ICT-Framework* zur Beschreibung von ICT-Skills finden sich in Engelhardt et al. (eingereicht), zur Itementwicklung und Validierung der im Projekt entwickelten ICT-Items in Engelhardt et al. (2017) und Engelhardt et al. (2019) sowie zur Skalierung der Items und Schätzung der ICT-Personenfähigkeiten in Wenzel et al. (2015) sowie Wenzel et al. (2016).

4.2.1 Das CavE-ICT-Framework: Gegenstandsbereich und interne Struktur von ICT-Skills

Zur Beschreibung von ICT-Skills wurde das Ziel verfolgt ein, trotz ständiger technologischer Weiterentwicklung und Veränderung, beständiges und dauerhaft nutzbares Framework zu entwickeln Um diesem Anspruch gerecht zu werden, wurde der Gegenstandsbereich von ICT-Skills nicht basierend auf den technischen, sich stetig verändernden Werkzeugen und Anwendungen strukturiert, sondern es wurde ein aufgabenzentrierter Ansatz basierend auf kognitiven Anforderungen verfolgt. ICT-spezifische Aufgaben und daraus resultierende Anforderungen bleiben bestehen, auch wenn sich die Anwendungen und Technologien weiterentwickeln oder durch moderne ersetzt werden. Das hat den Vorteil, dass auf Basis von ICT-Aufgaben entwickelte Testitems langlebiger und besser anpassbar sind, da sie unabhängig von den jeweiligen Technologien entworfen wurden (Wenzel et al., 2016). Des Weiteren kann so ein Testverfahren entwickelt werden, das sich an grundlegenden kognitiven Fertigkeiten und Fähigkeiten orientiert und nicht an der Oberfläche des Phänomens der Nutzung von ICT (in der Regel strukturiert durch verschiedene Technologien oder Applikationen) verhaftet ist.

Gegenstandsbereich von ICT-Skills. Für die Bearbeitung von vielen Aufgaben, die sich in ICT-Umgebungen stellen können, reicht allein der Besitz von Fertigkeiten im Umgang mit und das Wissen über Technologien nicht aus, um diese kompetent zu lösen. Andere generische Fertigkeiten, die nicht spezifisch mit ICT in Verbindung stehen, werden in vielen Konzeptionen zur Beschreibung von ICT-Skills trotzdem als implizit oder explizit relevant für die Lösung von Informationsproblemen in Computertechnologie-Umgebungen angesehen. Darunter fallen beispielsweise Lesen, Problemlösen, Rechnen, logisches Schlussfolgern, kritisches Denken und Metakognition (Calvani, Cartelli, Fini & Ranieri, 2008; International ICT Literacy Panel, 2002). Während beispielsweise die Anforderungen an numerische Fähigkeiten stark vom Aufgabeninhalt abhängen können, sind Problemlösefähigkeiten sowie Lesekompetenz und -verstehen für die erfolgreiche Bewältigung vieler ICT-spezifischer Aufgaben relevant (Engelhardt et al., eingereicht). Da in ICT-Umgebungen meist nicht nur die Verarbeitung von geschriebener Sprache, sondern auch von Bildmaterial, Film und Klang oder gesprochener Sprache, erforderlich ist, wird im Rahmen des CavE-ICT-Frameworks auf das Konstrukt der Medienrezeption zurückgegriffen. In diesem Sinne werden Problemlösen und Medienrezeption im Zusammenspiel mit Technikwissen als ICT-Skills verstanden, da diese Fertigkeiten eine Person erst dazu befähigen, ICT-Aufgaben zu lösen. Für eine tiefergehende Darstellung dieser drei Komponenten von ICT-Skills, angereichert durch empirische Befunde und unter Einbezug bestehender psychologischer Modelle und Theorien sei an dieser Stelle auf Engelhardt et al. (eingereicht) verwiesen. Die Zusammenhänge zwischen technischem Wissen, Problemlösen und Medienrezeption werden in Abbildung 4.1 dargestellt, wobei die den ICT-Skills zugeordneten Elemente grau markiert wurden.

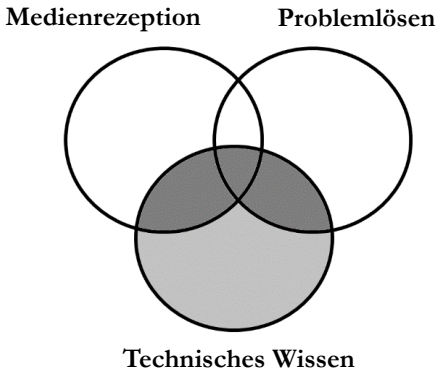

Abbildung 4.1 Gegenstandsbereich von ICT-Skills (aus Wenzel et al. 2016). Die den ICT-Skills zuzuordnenden Elemente sind grau (hell und dunkel) markiert.

Aus der Abbildung ist zu entnehmen, dass Problemlösen und Medienrezeption jeweils unter Nutzung von technischem Wissen ICT-Skills zugeordnet sind, wohingegen dies bei Problemlösen oder Medienrezeption ohne Rückgriff auf technisches Wissen nicht der Fall ist. Die Abbildung verdeutlicht auch, dass reines technisches Wissen ebenfalls den ICT-Skills zugeordnet wird. Die Größe der jeweiligen Schnittmengen soll allerdings keinen Rückschluss auf die relative Menge an alltäglich auftretenden ICT-spezifischen Aufgaben widerspiegeln, bei denen die jeweiligen Komponenten relevant sind. Im Folgenden soll der in Abbildung 4.1 eingefärbte Bereich genauer spezifiziert werden.

Interne Struktur von ICT-Skills. Bei der Frameworkentwicklung im Projekt CavE-ICT wurden vier Facetten beschrieben, um Aufgaben und erforderliche Fertigkeiten zur Lösung dieser Aufgaben zu charakterisieren und zu unterscheiden. Jede der postulierten Facetten ist in verschiedene Facettenebenen unterteilt. Eine Aufgabe kann durch Anforderungen auf vier Facettenebenen beschrieben werden: Sie erfordert einen speziellen **kognitiven Prozess**, ist innerhalb einer **sozialen Interaktion** sowie in einer bestimmten **Situation** zu lösen und erfordert die Verarbeitung einer bestimmten **Modalität** aufgrund der jeweiligen Inhaltsdarstellung.

Im Zentrum des CavE-ICT-Frameworks stehen fünf **kognitive Prozesse** (zugreifen, managen, integrieren, bewerten und erzeugen) die bei der Lösung von ICT-spezifischen Aufgaben relevant sind. Diese sind dem Framework des International ICT Literacy Panel (2002) entliehen, welches eines der einflussreichsten und bekanntesten Frameworks zur Strukturierung des Inhaltsbereichs von ICT-Skills darstellt (Ferrari et al., 2012). Das Framework des International ICT Literacy Panel hat zudem einen starken Einfluss sowie große Überlappung mit vielen weiteren Frameworks, die in den letzten Jahren entwickelt wurden (bspw. ICILS, Fraillon & Ainley, 2010; DigComp 2.1, Carretero, Vuorikari & Punie, 2017; Calvani et al., 2008; Eshet-Alkalai, 2004). *„Zugreifen"* beschreibt das Wissen darüber wie Informationen in ICT-Umgebungen abgerufen und gesammelt werden können. Eine ICT-spezifische Anforderung stellt dabei beispielsweise die oftmals große Vielfalt an Navigationsmöglichkeiten in ICT-Umgebungen dar, welche Desorientierung beim Nutzer auslösen kann. *„Managen"* bezieht sich darauf, bestehende Organisations- und Klassifikationsschemata anzuwenden. ICT-spezifische Anforderungen umfassen die Handhabung bekannter oder neuartiger, wenig bis hoch komplexer Applikationen zum Zwecke des Informationsmanagements (Calvani et al., 2008). *„Integrieren"* bezieht sich auf die zusammenfassende, vergleichende und/oder gegenüberstellende Darstellung und Interpretation von Informationen. Dabei erfordert die schiere Menge an über ICT zugänglichen Informationen, eine begründete Auswahl und selbstbestimmte Integration von Informationen aus verschiedensten Quellen (Metzger, 2007; Edmunds & Morris, 2000). Die Beurteilung der Qualität, Relevanz, Nützlichkeit oder Passung von Informationen wird unter *„Bewerten"* subsummiert. Um mit der großen Menge an über ICT zugänglichen Informationen umgehen zu können, ist die Beurteilung des Wertes von gefundenen Informationen besonders wichtig (Whittaker & Sidner, 1996). Dazu zählt auch die Beurteilung der Vertrauenswürdigkeit von Informationen, da beispielsweise bei Veröffentlichungen im Internet keine redaktionellen Kontrollmechanismen angenommen

werden können (Rieh, 2002; Lorenzen, 2001). Schließlich bezieht sich *„Erzeugen"* darauf wie Informationen durch Anpassung, Anwendung, Darstellung oder Verschriftlichung neu geschaffen werden können. Dabei stellt beispielsweise die angemessene Darstellung von Informationen unter Nutzung einer aus unzähligen, durch ICT eröffneten Bearbeitungsmöglichkeiten eine spezifische Anforderung an den Nutzer dar (Horz, Winter & Fries, 2009; Cox, Vasconcelos & Holdridge, 2010).

Der Aspekt der gemeinsamen Lösung von ICT-Aufgaben, die Nutzung von Technologien zum Teilen von Informationen und zur Kommunikation mit Anderen ist auch ein wichtiger Bestandteil verschiedener anderer ICT-Frameworks (ICILS, Fraillon & Ainley, 2010; DigComp 2.1, Carretero et al., 2017) und wird daher ebenfalls in der hier vorgestellten Konzeption im Rahmen des CavE-ICT-Projekts aufgegriffen. Die Facette der **sozialen Interaktion** bildet dementsprechend ab, ob eine Aufgabe in ICT-Umgebungen *„Individuell"* oder in der Kommunikation mit anderen *„Kollektiv"* angegangen wird.

Des Weiteren konnte gezeigt werden, dass sich die Computernutzung von Schülerinnen und Schülern zwischen kommunikationsbezogenen, eher in der Freizeit genutzten Anwendungen von eher arbeitsbezogenen Anwendungen unterscheidet (Senkbeil & Ihme, 2017). Auch das online-Leseverhalten von Schülerinnen und Schülern kann in freizeit- oder kommunikationsorientiertere und bildungs- oder berufsbezogene Nutzung von Texten unterschieden werden. Darüber hinaus kann davon ausgegangen werden, dass sich die Einstellungen von Personen zu Computern in Abhängigkeit vom jeweiligen Nutzungskontext unterscheiden (Naumann, Richter & Groeben, 2001; Richter et al., 2010). Die Facette der **Situation** berücksichtigt daher, dass ICT-Skills im täglichen Leben in verschiedenen Kontexten angewendet werden. Ob die Situation in der eine ICT-Aufgabe sich stellt *„Persönlich"*, *„Bildungsbezogen"* oder *„Beruflich"* ist kann mit

unterschiedlichen Erfahrungen in der Nutzung wie auch vorhandenen ICT-Skills verbunden sein.

Schließlich bezieht sich die Facette der **Modalität** auf verschiedene Repräsentationsformen in Informationen in ICT-Umgebungen und daraus resultierenden Anforderungen. Die Facettenebenen ergeben sich aus dem Integrativen Modell des Text- und Bildverstehens (Schnotz, 2005) und unterscheiden, ob Informationen *„Visuell"* oder *„Auditiv"* präsentiert werden, wobei diese Informationen jeweils konkret *„Analog"* oder *„Abstrakt"* sein können. Die Modalität wird als wichtiges Aufgabenmerkmal angesehen, welches Anforderungen von ICT-Aufgaben bestimmt, da beispielsweise gezeigt werden konnte, dass jüngere Personen grafische Benutzeroberflächen schneller dekodieren können als Ältere (Eshet-Alkalai & Amichai-Hamburger, 2004) und die Integration von Texten und Diagrammen aufgrund eines „Split-attention-Effekts" (bspw. Horz, 2009) zu einer zusätzlichen kognitiven Belastung führen kann (Chandler & Sweller, 1991; Horz & Schnotz, 2010).

In Abbildung 4.2 wird die interne Struktur von ICT-Skills im CavE-ICT-Framework durch die soeben erläuterten vier Facetten und jeweiligen Facettenebenen schematisch dargestellt.

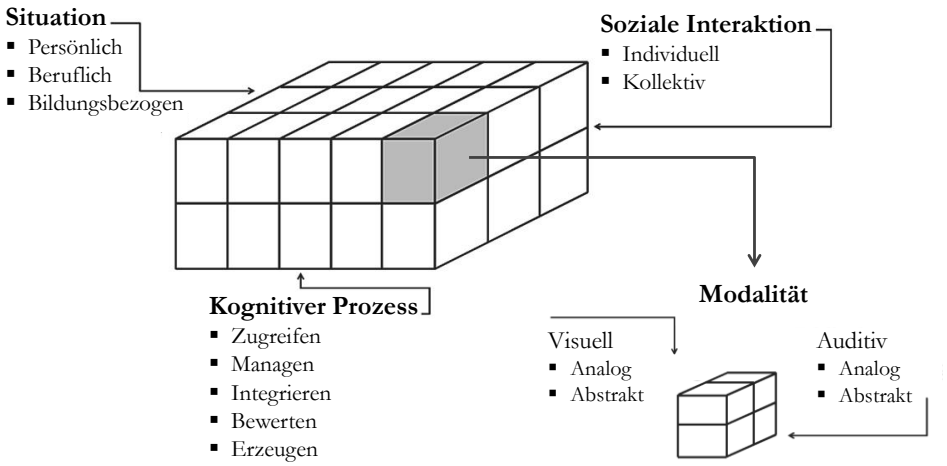

Abbildung 4.2 Interne Struktur von ICT-Skills im CavE-ICT-Framework (in Anlehnung an Wenzel et al., 2016).

Jeder Würfel in der internen Struktur repräsentiert mögliche ICT-Aufgaben und daraus resultierende Aufgabenmerkmale und Lösungsanforderungen. Um diesen Anforderungen zu begegnen, werden spezifische Fertigkeiten benötigt. Durch die Definition der zuvor dargestellten vier Facetten zur Beschreibung der internen Struktur des ICT-Skills-Konstrukts wurde eine möglichst vollständige und hoch strukturierte Beschreibung von ICT-Skills realisiert. Dadurch wird die Vielfalt möglicher ICT-Aufgaben dargestellt und die Art von möglichen ICT-Items, die zur Messung des Konstrukts erstellt werden können, dargelegt.

4.2.2 Itementwicklung

Auf Grundlage des vorgestellten und in Abbildung 4.2 dargestellten theoretischen Frameworks wurde abgeleitet, welche Art von Items die konstruktbestimmenden Fertigkeiten repräsentieren. Allerdings wurde bei der Konstruktion der Items auf eine Unterscheidung zwischen konkret analogen und abstrakten Tönen verzichtet, da die Darbietung auditiver Stimuli für die Erhebungen im Projektkontext nicht realisierbar war. Zudem wurde reines technisches Wissen ohne Medienrezeption und/oder Problemlösen

in den zu entwickelnden Items nicht vordergründig abgebildet, da dieses eher in einfachen Routineaufgaben repräsentiert wird. Eine Zielsetzung der Itementwicklung war es, mit den entwickelten Testaufgaben das Assessmentframework möglichst umfassend und balanciert abzubilden. Dementsprechend wurden Items entwickelt, die Kombinationen der Facettenebenen operationalisieren. Die computerbasierte Umsetzung der ICT-Items wurde unter Verwendung des Autorentools CBA Itembuilder (Rölke, 2012) realisiert. Um einheitliche Designs für wiederkehrende Elemente in den Items festzulegen, wurden zunächst Vorlagen entwickelt und durchgängig genutzt. Dabei wurde auf eine ausreichende Abstraktion von gängigen Applikationen geachtet, um keine Vorteile für bestimmte Nutzergruppen aufgrund des Designs zu bieten. Zudem wurde angestrebt eine möglichst große Bandbreite gängiger Computerapplikationen abzubilden, so dass in den entwickelten Items ICT-spezifische Aufgaben in Browser, E-Mail, Ordnerstruktur, Text-, Präsentations- und Tabellenkalkulationssoftware repräsentiert sind. Die simulationsbasierte Umsetzung dieser Umgebungen in den ICT-Items sollte eine möglichst nah an alltäglichen ICT-Nutzungen orientierte Repräsentation des zu messenden Merkmals gewährleisten.

Im Zuge der Itementwicklung wurde zudem direkt eine automatische Bewertung von korrekten und inkorrekten Lösungen wie auch Teillösungen für spätere Itemanalysen implementiert. Die ICT-Items wurden dementsprechend bei korrekter Lösung mit 1 und bei inkorrekter Lösung mit 0 bewertet. Wurde bei einem Item keine Interaktion vorgenommen und direkt zum nächsten Item weitergegangen, konnte dieses automatisch als ausgelassen identifiziert werden (engl. Omitted Response). Items, die der Testteilnehmende zur Bearbeitung nicht erreichte (engl. Not Reached), weil beispielsweise die Testzeit nicht zur Bearbeitung aller Items ausgereicht hat, konnten so ebenfalls identifiziert werden.

Eine ausführlichere Beschreibung des Itementwicklungsprozesses findet sich in Wenzel et al. (2015) und Wenzel et al. (2016). Zudem können Anhang A zwei Beispielitems entnommen werden (Abbildung A.1 und Abbildung A.2).

Letztlich lag ein Pool von 70 Items vor, der im Zuge des CavE-ICT-Feldtests unter Nutzung von IRT-Modellen kalibriert wurde. In Tabelle 4.1 werden die Itemzahlen nach Facetten und Facettenebenen aufgeschlüsselt dargestellt.

Tabelle 4.1
Itemzahlen nach Facetten und Facettenebenen des Frameworks zur Messung von ICT-Skills (aus Wenzel et al., 2016)

	Soziale Interaktion						
	Individuell			Kollektiv			
	Situation						
Kognitive Prozesse	Persönlich	Beruflich	Bildungs-bezogen	Persönlich	Beruflich	Bildungs-bezogen	Summe
Zugreifen	3	1	5	2	2	0	13
Managen	4	4	1	2	6	7	24
Integrieren	1	1	1	2	3	2	10
Bewerten	5	1	4	1	2	0	13
Erzeugen	0	3	2	1	1	3	10
Summe	13	10	13	8	14	12	70

4.3 Der CavE-ICT-Feldtest

Mit dem Ziel, die entwickelten ICT-Items zu kalibrieren und im nächsten Schritt die ICT-Skala zu validieren, wurde ein Feldtest durchgeführt. Dieser Feldtest kann somit im Sinne der Ausführungen in Kapitel 2.3 auch als Kalibrierungsstudie im Zuge der Testentwicklung bezeichnet werden.

Die geplante Testzeit lag bei insgesamt drei Schulstunden und entsprechend 135 Minuten, wobei jeweils 60 Minuten für die Vorgabe der ICT-Items und 60 Minuten für die Erhebung von Validierungsinstrumenten vorgesehen wurden. Zwischen beiden

Testteilen wurde eine 10-minütigen Pause gemacht. Auf die Validierung wird im Folgenden nicht näher eingegangen, da diese nicht im Zentrum der vorliegenden Arbeit stehen. Nähere Informationen zum Validierungsdesign, den eingesetzten Instrumenten und Befunden dazu finden sich in Engelhardt et al. (2019).

Zur Kalibrierung der 70 ICT-Items wurde ein balanciertes unvollständiges Testheftdesign (Youden-Square-Desgin; Frey et al., 2009; genutzte Software: Youden 1.0, Frey & Annageldyev, 2015) eingesetzt. Dabei wurde den Schülerinnen und Schülern jeweils nur einen Teil der ICT-Items zur Bearbeitung vorgelegt. Bei einer verfügbaren Testzeit von 60 Minuten, abzüglich eines etwa 10-minütigen zu Beginn der Testung eingebundenen Tutorials, das mit der Testumgebung vertraut machen sollte, wurden den Schülerinnen und Schülern zwischen 29 und 35 ICT-Items vorgelegt. Dabei wurde eine durchschnittliche Bearbeitungszeit je Item von einer Minute und 30 Sekunden kalkuliert. Weitere Einzelheiten zur Testdurchführung finden sich bei Wenzel et al. (2016).

Stichprobe. Im Zuge des CavE-ICT-Feldtests wurde in insgesamt 33 Schulen, in Baden-Württemberg und Rheinland-Pfalz erhoben, wobei alle Schulformen (Gymnasien, Realschulen, Hauptschulen, Gesamtschulen) vertreten waren. Es konnten Angaben von 983 Schülerinnen und Schülern gesammelt werden. 71.0 % der Befragten besuchten die 9. Klasse, 26.3 % gingen zum Zeitpunkt der Testung in die 10. Klasse. Die Schüler waren im Mittel 15.21 Jahre alt (SD = 0.57). 51.0 % der Testteilnehmer waren männlich, 46.4 % weiblich (2.6 % machten keine Angabe zum Geschlecht). ICT-Items wurden von 766 Schülerinnen und Schülern bearbeitet, da ein Teil der ICT-Items zu Validierungszwecken verändert und nicht zusammen mit den eigentlich zu kalibrierenden ICT-Items vorgegeben werden konnte (Engelhardt et al., 2019). Im Mittel lagen etwa 287 Antworten pro Item vor mit einer Range von 245 bis 325 Antworten pro ICT-Item.

4.3.1 Itemkalibrierung und -selektion

Viele Leistungstests, vor allem im Kontext von LSAs, verfolgen das Ziel Fähigkeiten zu messen, die sich gemäß ihrer theoretischen Konzeption aus anderen, spezielleren Subfähigkeiten zusammensetzen (Brandt, 2015). Ein gängiger Ansatz bei der Berechnung aller gewünschten Fähigkeitsschätzungen ist es, die Daten zum einen unter Nutzung eines eindimensionalen Modells (zur Kalibrierung der Globalskala) und zum anderen mit Hilfe eines mehrdimensionalen Modells (zur Kalibrierung der Subskalen) zu analysieren. Dieses Vorgehen wurde auch im CavE-ICT-Projekt gewählt. Entsprechend des zuvor dargestellten Frameworks zur Beschreibung von ICT-Skills wurden die fünf kognitiven Prozesse als zentraler Aspekt zur Differenzierung von ICT-spezifischen Aufgaben angesehen. Daher sollten nicht nur Aussagen über generelle ICT-Skills generiert werden, sondern auch spezifische Fähigkeiten hinsichtlich der einzelnen kognitiven Prozesse rückgemeldet werden können. Daher wurde im Zuge der Itemselektion zunächst ein eindimensionales Rasch-Modell genutzt, um die 70 ICT-Items zu untersuchen. Eines der Items wurde allerdings nie korrekt gelöst, woraufhin es zunächst hinsichtlich technischer Funktionalität, Instruktionsgenauigkeit und Bewertung geprüft wurde. Auf Basis dieser Überprüfung konnte das Item sinnvoll umkodiert und in die Analysen einbezogen werden. Die Modellparameter wurden mit der Software ConQuest (Wu et al., 2007) geschätzt. Mit dem Ziel einen Itempool zusammenzustellen, der eine homogene Erfassung von ICT-Skills ermöglicht, wurden die ICT-Items zunächst hinsichtlich ihrer psychometrischen Güte geprüft, aber auch inhaltliche Aspekte zur Beurteilung herangezogen. Neben der Itemschwierigkeit wurden der Itemfit, anhand des *WMNSQ* und entsprechenden *t*-Wertes (siehe Kapitel 2.3) und die punkt-biseriale Korrelation des Itemtestwertes mit dem Testwert der Gesamtskala sowie differentielle Effekte (Differential-Item-Functioning, DIF; Osterlind & Everson, 2009) bezüglich des Geschlechts (Gender DIF) als Kriterien für die psychometrische Güte berücksichtigt. Der Gender DIF wurde mit Hilfe eines

Multifacetten-Raschmodells (Linacre, 1994) unter Kontrolle des Haupteffektes Geschlecht berechnet. Für die abschließende Skalierung wurden schließlich 64 der 70 ICT-Items selektiert. Bei zwei der sechs ausgeschlossenen Items wurden sehr niedrige punktbiseriale Korrelationen beobachtet. Insgesamt vier Items zeigten in der psychometrischen und inhaltlichen Analyse Hinweise auf Gender DIF. Eines dieser Items wies zudem eine zu niedrige Modellpassung auf.

Mit Blick auf die über Personen gemittelten Bearbeitungszeiten der 64 selektierten ICT-Items zeigte sich eine große Heterogenität. Die Spannbreite der Itembearbeitungszeiten lag zwischen etwa 40 Sekunden und etwas über vier Minuten mit einem Mittelwert von einer Minute und 45 Sekunden (*SD* = 41 Sekunden), was über der im Zuge der Zusammenstellung von Testheften antizipierten Zeit von einer Minute und 30 Sekunden liegt.

4.3.2 Abschließende Schätzung von Item- und Personenparametern

Für die abschließende Skalierung mit Fähigkeitsschätzungen für die kognitiven Prozesse wurden die eindimensional ermittelten Itemparameter fixiert und ein fünf-dimensionales Raschmodell angewendet. Die ICT-Itemschwierigkeiten nehmen Werte zwischen -2.96 und 4.24 Logits mit einem Mittelwert von *M* = 0.34 Logits und einer Standardabweichung von *SD* = 1.58 Logits an. In Tabelle 4.2 werden die Itemzahlen und Itemschwierigkeiten, nach kognitivem Prozess unterteilt, abgebildet. Zudem stellt Abbildung 4.3 die Verteilung der Itemschwierigkeiten zusammen mit der geschätzten ICT-Fähigkeitsverteilung dar. Eine Auflistung aller selektierten CavE-ICT-Items mit Zuordnung zu den in der Rahmenkonzeption definierten Facettenebenen, dargestellten Applikationen, Itemkennwerten und durchschnittlichen Bearbeitungszeiten kann Anhang A, Tabelle A.1 entnommen werden.

Tabelle 4.2

Itemanzahl sowie Minimum (MIN), Maximum (MAX), Mittelwert (M) und Standardabweichung (SD) der Itemschwierigkeiten für die ICT-Gesamtskala sowie nach kognitiven Prozessen aufgeschlüsselt (aus Wenzel et al., 2016)

Kognitive Prozesse	Itemanzahl	Itemschwierigkeiten			
		MIN	*MAX*	*M*	*SD*
Zugreifen	13	-1.390	3.467	0.659	1.476
Managen	20	-2.327	4.243	0.544	1.708
Integrieren	10	-0.765	3.541	0.912	1.456
Bewerten	11	-2.043	1.674	-0.259	1.222
Erzeugen	10	-2.961	2.517	-0.388	1.729
ICT-Gesamt	64	-2.961	4.243	0.341	1.580

Die Berechnung der Personenparameter erfolgte unter Nutzung des Expected-A-posteriori-Schätzers (EAP; Bock & Aitken, 1981, Bock & Mislevy, 1982). Die Varianz der geschätzten Personenparameter liegt bei 0.33 (entspricht einer Standardabweichung von 0.57), wobei der Mittelwert der Fähigkeitsverteilung auf Null fixiert wurde (Wenzel et al., 2016). Bei Betrachtung der ICT-Items, dargestellt in Abbildung 4.3, zeigt sich, dass die für die Skalierung ausgewählten Items den Merkmalsbereich insgesamt gut abdecken.

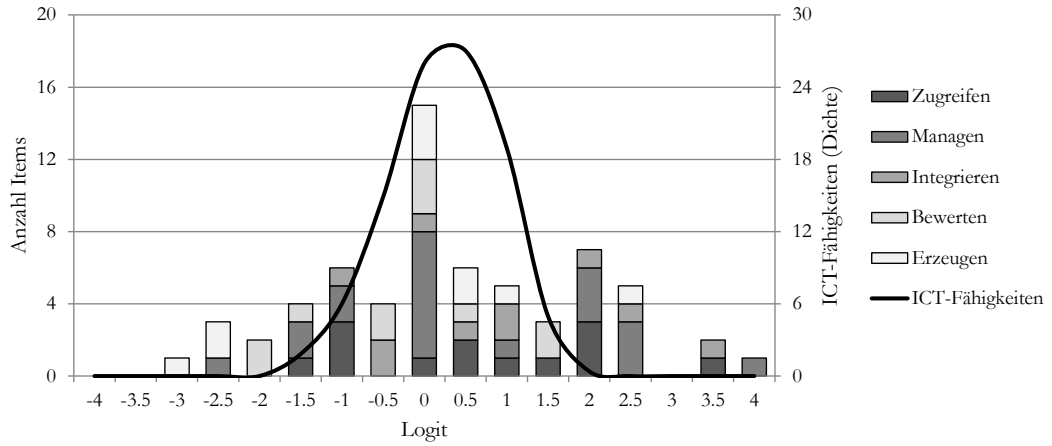

Abbildung 4.3 Itemzahlen nach Itemschwierigkeiten und Personenparameter (ICT-Fähigkeiten) der für die abschließende Skalierung selektierten ICT-Items (aus Wenzel et al., 2016).

Eine gute Passung zwischen Item- und Personenparametern ist im Hinblick auf hohe Differenzierungsfähigkeit bei hoher Messeffizienz erstrebenswert. Es wird aber auch deutlich, dass eine gewisse Anzahl von Items höhere Itemschwierigkeiten aufweisen, und damit etwas über der Verteilung der Personenfähigkeiten in der CavE-ICT-Stichprobe liegen. Abbildung 4.4 zeigt zudem die ICT-Items und die Fähigkeitsverteilung nach kognitiven Prozessen.

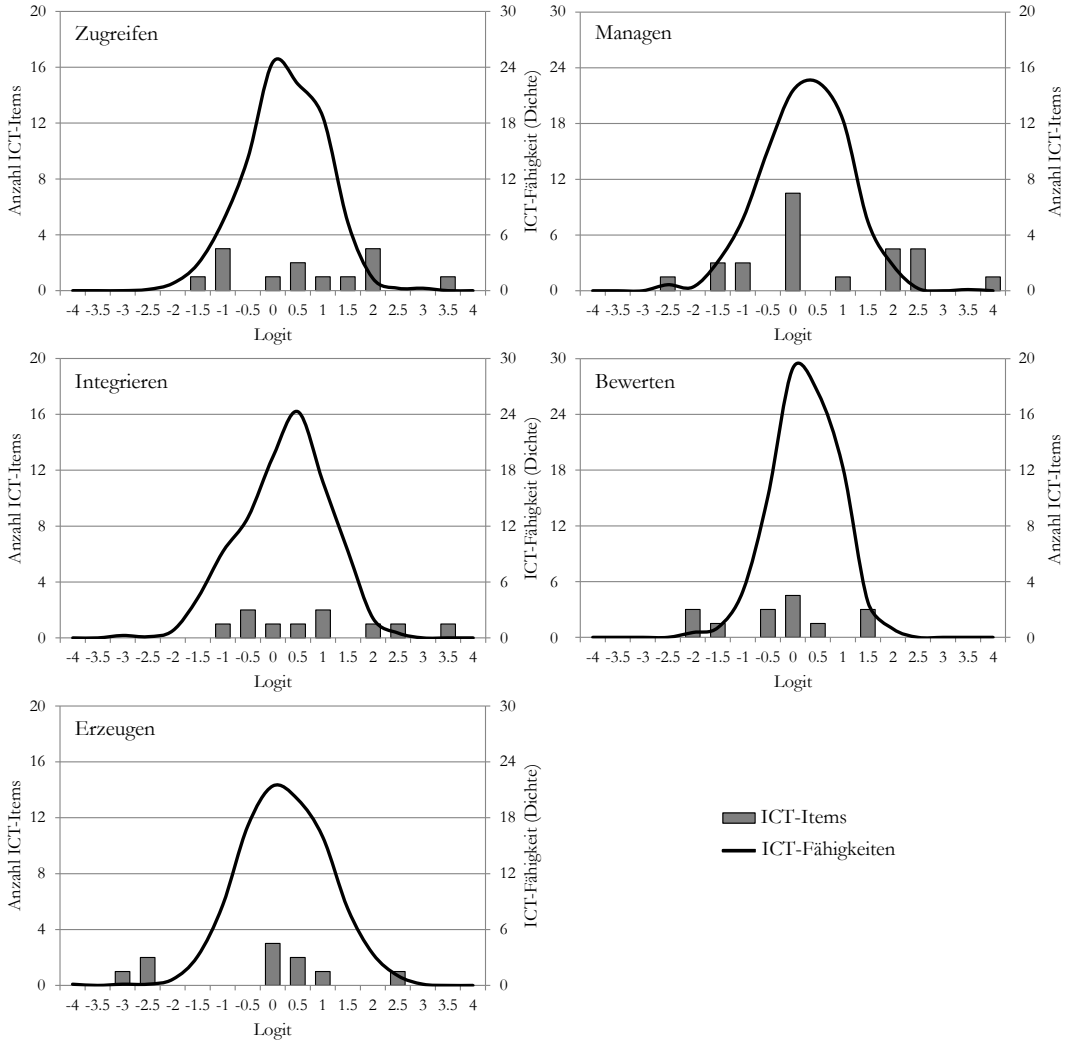

Abbildung 4.4 Itemzahlen nach Itemschwierigkeit und Personenparameter (ICT-Fähigkeiten) der für die abschließende Skalierung selektierten ICT-Items aller kognitiver Prozesse (Subskalen).

Es ist zu erkennen, dass beispielsweise die Verteilung der Items für den kognitiven Prozess „Erzeugen“ schwerpunktmäßig im sehr niedrigen und mittleren Bereich als bimodal beschrieben werden kann. Für die kognitiven Prozesse „Zugreifen“ und „Integrieren“ zeigen sich mehr Items im höheren Schwierigkeitsbereich. Die Itemverteilungen für die Prozesse „Managen“ und „Bewerten“ sind relativ ausgewogen, wobei für „Managen“ eine recht hohe Anzahl von Items (10) im Bereich von -.25 bis .25 Logits liegen.

Für die ICT-Gesamtskala ergibt sich ein Reliabilitätskoeffizient von .796 (EAP/PV-Reliabilität). Die Reliabilitäten der kognitiven Subskalen liegen zwischen .639 und .727 (Zugreifen: .727, Managen: .718, Integrieren: .639, Bewerten: .639, Erzeugen: .700). Die latenten Korrelationen zwischen den verschiedenen kognitiven Prozessen sind zusammen mit den Kovarianzen und Varianzen der Personenparameterschätzungen in Tabelle 4.3 abgetragen. Die Zusammenhänge zwischen den kognitiven Prozessen bewegen sich im Bereich von .476 bis .718. Zudem weisen die Anteile erklärter Varianz an Gesamtvarianz der Facettenebenen mit 22.66 % bis 51.55 % darauf hin, dass die kognitiven Prozesse zwar Gemeinsamkeiten aufweisen, darüber hinaus aber auch spezifische Inhalte abbilden (Wenzel et al., 2016).

Tabelle 4.3

Latente Korrelationen (unterhalb der Diagonalen), Kovarianzen (oberhalb der Diagonalen) und Varianzen der Personenparameter auf der Logit-Skala

	Kognitive Prozesse				
	Zugreifen	Managen	Integrieren	Bewerten	Erzeugen
Zugreifen	-	0.341	0.301	0.186	0.244
Managen	.718	-	0.409	0.257	0.333
Integrieren	.591	.717	-	0.248	0.329
Bewerten	.476	.586	.528	-	0.219
Erzeugen	.514	.625	.576	.499	-
Varianz	0.424	0.532	0.613	0.362	0.532

4.3.3 CAT Simulation

Auf Basis der Skalierungsergebnisse liegt ein ICT-Itempool vor, der für die Erprobung eines CAT-Algorithmus geeignet ist. Die Itemparameter wie auch Eigenschaften der Zielpopulation wurden im Rahmen der Kalibringung geschätzt und konnten bei der Simulation eines eindimensionalen adaptiven Algorithmus verwendet werden (Wenzel et al., 2015). So wurde zur Generierung von Personenparametern die im Rahmen des CavE-ICT-Feldtests ermittelte Standardabweichung der geschätzten Personenparameter von 0.57 genutzt. Für den adaptiven Algorithmus wurde als Content-Management-Methode der Maximum-Priority-Index (MPI; Y. Cheng & Chang, 2009; siehe Kapitel 3.1.4) implementiert, um eine gleichverteilte Vorgabe von Items der unterschiedlichen kognitiven Prozesse, welche bei der Bearbeitung ICT-bezogener Aufgaben relevant sein können, zu forcieren. Der Test sollte beendet werden, wenn keine Items zur Vorgabe mehr verfügbar sind. Durchgeführt wurde die Simulationsstudie mithilfe der Software Multidimensional Adaptive Testing Environment (MATE Version 1.1.1; Kröhne & Frey, 2014) die zur Administration aber auch zur Simulation computerisierter adaptiver Tests genutzt werden kann. In Abbildung 4.5 werden die im CAT-Verlauf ermittelten Reliabilitäten als quadrierte Korrelation wahrer und geschätzter Fähigkeiten abgetragen.

Nach 17 vorgegebenen ICT-Items kann eine Reliabilität von .55 erreicht werden, was im Folgenden für den Bericht von Testwerten im Kontext von LSAs als ausreichend angesehen wird (siehe Kapitel 6). Eine für die Untersuchungen kleinerer Gruppen oder bei individualdiagnostischen Zwecken akzeptable Reliabilität von .70 (Danner, 2015; siehe Kapitel 8) wird erst mit der Administration von 33 ICT-Items erreicht.

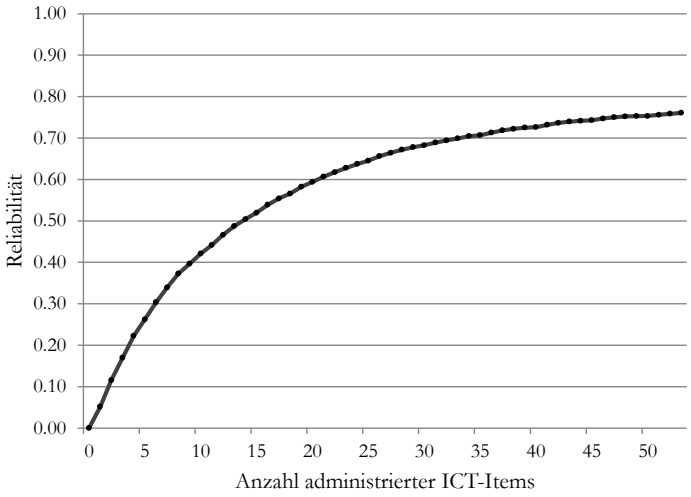

Abbildung 4.5 Reliabilität in Abhängigkeit der Anzahl administrierter ICT-Items im Verlauf des eindimensionalen adaptiven Tests

Legt man die im CavE-ICT-Feldtest beobachtete mittlere Bearbeitungszeit der 64 ICT-Items von einer Minute und 45 Sekunden zugrunde ist daher mit einer reinen Testzeit (ohne Tutorial) von circa 30 Minuten bei Vorgabe von 17 Items beziehungsweise 58 Minuten bei Vorgabe von 33 ICT-Items zu rechnen.

Ziel dieses Kapitels war es die, den Studien dieser Arbeit zu Grunde gelegten theoretischen Annahmen zur Beschreibung des Konstrukts ICT-Skills darzustellen sowie den CavE-ICT-Feldtest und die im Zuge dessen gewonnen Daten, die in den Studien II und III (Kapitel 7 und 8) genutzt werden vorzustellen. Über die Studien dieser Arbeit sollen Möglichkeiten exploriert werden wie die im Projekt CavE-ICT entwickelten Items in andere effiziente Testinstrumente (andere als die im Projekt vorgestellte 64 Items umfassende ICT-Gesamtskala oder den eindimensionalen adaptiven Algorithmus) überführt werden können.

5 Problemstellungen, Zielsetzungen und Fragestellungen

Im Rahmen der vorliegenden Arbeit wird untersucht, auf welche Weise komplexe Konstrukte wie ICT-Skills (siehe Kapitel 4) effizient, zuverlässig und psychometrisch adäquat unter Einsatz zeitgemäßer Methoden erfasst werden können. Dabei steht die Testentwicklung und -zusammenstellung im Fokus der Arbeit. Verschiedene Autoren benennen Schritte oder Phasen der Testentwicklung (bspw. Downing, 2006; Eid & Schmidt, 2014; Schmeiser & Welch, 2006). Zusammenfassend lassen sich fünf Kernschritte oder Stufen der Entwicklung von Testinstrumenten identifizieren, welche sowohl für die Zusammenstellung traditioneller wie auch IRT-basierter Tests (Kapitel 2) anwendbar sind. Am Anfang jeder Testentwicklung sollten stets Überlegungen zum Testinhalt und den anvisierten Testzielen stehen, aus denen sich wiederum Implikationen für das Itempooldesign ergeben. Anschließend gilt es, geeignete Items zu erstellen, um einen Pool zu generieren, der den zuvor angestellten Überlegungen entspricht. Die Itementwicklung wird meist ins Zentrum der Testentwicklung gerückt, da die Güte des Testinstruments elementar durch die Qualität des Itempools bestimmt wird. Die entwickelten Items sollten in einem nächsten Schritt evaluiert, beispielsweise durch Experten geprüft und im Zuge von Pilotierungsstudien analysiert werden. Als ungeeignet identifizierte Items sollten aus dem Pool eliminiert werden. Unter Nutzung des so geprüften Itempools erfolgt anschließend die Testzusammenstellung. Der entwickelte Test wird nachfolgend administriert und evaluiert, worauf folgend Berichte und Dokumentationen zum Test und dem Testentwicklungsprozess angelegt werden sollten. Auch die weitere Pflege (bspw. regelmäßige Prüfung der Aktualität von Testinhalten oder konkreter der Stabilität von Itemeigenschaften) des erarbeiteten Itempools wird als nachgelagerter Schritt im Zuge der Testentwicklung oftmals einbezogen. Eine ähnliche Abfolge von Schritten der Testentwicklung speziell für CAT stellen Thompson und Weiss (2011) vor. Sie empfehlen im ersten Schritt die Anwendung von Simulationsstudien, um

zu prüfen, ob und in welcher Form adaptives Testen im angestrebten Kontext sinnvoll sein kann. Dazu gehört die Abwägung, welche Vorteile es im jeweiligen Fall bereithält, aber auch welcher Mehraufwand gegebenenfalls zunächst geleistet werden muss. Auf Basis der so gewonnen Erkenntnisse sollte ein entsprechender Itempool entwickelt und kalibriert werden. Um den Itempool für CAT nutzen zu können, schlagen sie in einem vierten Schritt nicht nur die Festsetzung von Testspezifikationen vor, sondern empfehlen weitere Simulationen unter Nutzung der im Zuge der Kalibrierung erhobenen Daten (Hybrid-Simulationen). Dadurch können Testspezifikationen bereits im Vorfeld der Testadministration auf ihre Angemessenheit hin geprüft werden. Erst danach wird der Test beziehungsweise der CAT-Algorithmus eingesetzt. Die weitere Pflege des Itempools wird von Thompson und Weiss (2011) zwar nicht explizit als Schritt der Testentwicklung genannt, allerdings als eine wichtige weiterführende Aufgabe hervorgehoben. In der vorliegenden Arbeit werden drei Studien vorgestellt, die auf unterschiedliche Weise auf verschiedene Schritte der Testentwicklung fokussieren. Dabei wird sich eng an den Erfordernissen des konkreten Inhaltsbereichs, hier speziell an den Anforderungen der Messung von ICT-Skills orientiert. Während in den ersten beiden Studien im Rahmen dieser Arbeit der Frage nachgegangen wird, wie adaptive Testalgorithmen zur Erfassung von ICT-Skills eingesetzt werden können, beschäftigt sich die dritte Studie mit der Zusammenstellung eines verkürzten linearen ICT-Skills-Messinstruments. In Tabelle 5.1 werden die zuvor genannten Kernschritte der Testentwicklung als Zusammenfassung der Vorschläge von Downing (2006), Eid und Schmidt (2014) sowie Schmeiser und Welch (2006) mit denen von Thompson und Weiss (2011) speziell zur Entwicklung computerisierter adaptiver Tests vergleichend dargestellt. Zudem werden die Studien dieser Arbeit mit den Schritten der Testentwicklung in Bezug gesetzt.

Tabelle 5.1

Vergleich vorgeschlagener Schritte der Testentwicklung und Bezug zu den Studien der vorliegenden Arbeit

Schritte der (IRT-basierten) Testentwicklung (Downing, 2006; Eid & Schmidt, 2014; Schmeiser & Welch, 2006)	Schritte der Entwicklung von CATs (Thompson & Weiss, 2011)	Ansatz der einzelnen Studien im Rahmen dieser Arbeit
Überlegungen zu Testinhalten, -zielen und -spezifikationen und zum Itempooldesign	Simulationsstudien zur Klärung der Anforderungen an Test- und Itempooldesign	**Studie I:** Simulationsstudie vor der Itementwicklung zur Spezifikation von Itempoolgröße und Testlänge eines CAT zur Messung von ICT-Skills
Entwicklung des Itempools	Entwicklung des Itempools	
Itemkalibrierung/ -analyse/-evaluation	Itemkalibrierung/ -analyse/-evaluation	**Studie III:** Zusammenstellung von Kurztests zur Erfassung von ICT-Skills unter Nutzung der kalibrierten CavE-ICT-Items
Testzusammenstellung	Festlegung von Testspezifikationen durch Erprobung in Simulationsstudien	**Studie II:** Simulationsstudie zur Erprobung verschiedener CAT-Algorithmen unter Nutzung der CavE-ICT-Feldtestdaten; **Studie III:** Simulationsstudie zur Evaluation der zusammengestellten ICT-Kurztests
Testadministration und Evaluation	Testveröffentlichung/ -administration	
Berichtlegung, Dokumentation und Pflege des Itempools	Pflege des Itempools	

Alle drei Studien beziehen sich auf Annahmen und Arbeitsergebnisse des in Kapitel 4 vorgestellten Projekts CavE-ICT. Es wird daher auf die im Projekt erarbeitete, theoretische Rahmenkonzeption zur Beschreibung von ICT-Skills Bezug genommen und es werden Daten des CavE-ICT-Feldtests sowie Ergebnisse aus der Itemkalibrierung genutzt. Die erarbeiteten Erkenntnisse können insbesondere für die Erfassung von mehrdimensionalen Konstrukten oder facettierten Merkmalen in LSAs, aber auch darüber hinaus für gruppen- und individualdiagnostische Zwecke genutzt werden. Die konkreten Problemstellungen, die Zielsetzungen und die daraus abgeleiteten Fragestellungen der drei empirischen Studien der vorliegenden Arbeit werden in den folgenden Unterkapiteln detailliert vorgestellt.

5.1 Studie I – Spezifikation von Itempoolgröße und Testlänge eines computerisierten adaptiven Tests zur multidimensionalen Messung von ICT-Skills

Zur verhaltensnahen Erfassung von ICT-Skills wurde im Rahmen des Forschungsprojekts CavE-ICT ein Itempool entwickelt, mit dem auch die Umsetzung von CATs ermöglicht wird. Da die Konstruktion von simulationsbasierten und damit verhaltensnahen ICT-Items sehr aufwändig und komplex ist, sollte vorab über eine Simulationsstudie untersucht werden, welche Größe der zu entwickelnde Itempool haben muss, um für verschiedene Testlängen hinreichend präzise Testwerte für die einzelnen Teilskalen gewährleisten zu können. Dabei sollte auch geprüft werden, welche verschiedenen adaptiven Testalgorithmen zum Einsatz kommen könnten, um die angenommene multidimensionale Struktur des Konstrukts von ICT-Skills in Bezug auf die kognitiven Prozesse (Kapitel 4) abzubilden.

Bei der Erfassung mehrerer korrelierter Merkmalsdimensionen können multidimensionale Fähigkeitsschätzer (Bloxom & Vale, 1987; Brandt & Duckor, 2013) und MAT (Luecht, 1996; Segall, 1996) eingesetzt werden, um Messpräzision und Messeffizienz (Frey & Seitz, 2009; W.-C. Wang, Chen & Cheng, 2004) zu steigern (siehe auch Kapitel 3.2). Dabei werden zum einen Korrelationen zwischen den Merkmalsdimensionen im Zuge der Itemauswahl und für die abschließende Fähigkeitsschätzung genutzt. Zum anderen kann eine effizientere Itemselektion durch die Verfügbarkeit eines multidimensionalen Itempools erreicht werden (Li & Schafer, 2005). Allerdings richtet sich beim MAT die Itemauswahl häufig ausschließlich nach einem statistischen Optimalitätskriterium (Kapitel 3.1.3), sodass sich Items verschiedener Dimensionen im Testverlauf abwechseln. Aus psychologischer Perspektive kann eine sequentielle Vorgabe von Items für eine Dimension nach der anderen aber durchaus wünschenswert sein. Da neben Itemkontexteffekten (Yousfi & Bohme, 2012) durch den

Wechsel von Items verschiedener Dimensionen höhere kognitive Anforderungen an die Testbearbeitung resultieren und so auch metakognitive oder motivationale Effekte auf die Probanden gezeitigt werden können. Im Hinblick auf Effekte, welche die Motivation zur Testbearbeitung betreffen, lässt sich vermuten, dass ein Zusammenwirken verschiedener Test-, Personen- und Situationsmerkmale vorliegt (Frey & Ehmke, 2007). Dabei könnte auch eine höhere Heterogenität der vorgelegten Items beim MAT eine Rolle spielen. Solche Effekte könnten auch bei der Erfassung von ICT-Skills auftreten, sodass die dimensionsunspezifische Vorgabe von Items verwirrend auf die Testperson wirken kann. Um derartige Übertragungseffekte zu vermeiden, sollte möglichst bereits im Vorfeld der Itementwicklung geprüft werden, ob CAT-Algorithmen, die eine sequentielle Vorgabe von Items je Dimension ermöglichen, eine sinnvolle Alternative zum MAT darstellen. Dass solche Testalgorithmen unter bestimmten Umständen Alternativen zu MAT sein können, haben Kröhne, Goldhammer und Partchev (2014) in einer Simulationsstudie mit Echtdaten gezeigt. Sie argumentieren, dass bereits bestehende unidimensionale Tests für korrelierte Merkmalsausprägungen gemeinsam unter Nutzung multidimensionaler Algorithmen auch sequentiell administriert werden könnten, um die Messeffizienz zu steigern. Weitere Studien in diesem Zusammenhang liegen bisher nicht vor, sodass das Potenzial sequentieller dimensionsspezifischer Itemvorgabe in multidimensionalen Testalgorithmen bislang noch kaum erforscht ist.

Durch die vorliegende Studie I sollen diese Testspezifikationen, aber auch Anforderungen an das ICT-Itempooldesgin, auf Basis von Simulationen exploriert werden, was unter den zuvor dargestellten Schritt 1 der IRT- beziehungsweise computerisierten adaptiven Testentwicklung fällt. Die in Studie I aufgeworfenen Fragestellungen lassen sich in zwei Aspekte unterteilen: Zunächst sollen Hinweise zum zu entwickelnden ICT-Itempool abgeleitet werden. Dazu wurde die Größe des zur

Verfügung stehenden Itempools variiert, wobei eine Gleichverteilung von Items auf die fünf Dimensionen angenommen wird. Daraus ergibt sich folgende Frage:

> Forschungsfrage 1.1: Wie groß sollte der zu entwickelnde ICT-Itempool sein, um eine präzise, effiziente und zuverlässige Messung von ICT-Skills zu ermöglichen?

Darüber hinaus sollen vorab bereits mögliche, relevante Spezifikationen zur Testzusammenstellung und -administration geprüft werden. Daher soll zunächst geklärt werden, wie lang der auf Basis des zu entwickelnden Itempools administrierte adaptive Test mindestens sein müsste.

> Forschungsfrage 1.2: Wie viele Items sollten im späteren adaptiven ICT-Test vorgelegt werden, um eine präzise, effiziente und zuverlässige Messung von ICT-Skills zu ermöglichen?

Dabei soll auch die Effizienz verschiedener Testalgorithmen, die die angenommene multidimensionale Struktur des ICT-Skills-Konstrukts in unterschiedlicher Weise berücksichtigen, eingehend untersucht werden. Ein MAT-Algorithmus mit multidimensionaler vorläufiger und abschließender Personenparameterschätzung und drei Varianten von eingeschränkten MAT-Algorithmen, welche jeweils dimensionsspezifische eindimensionale vorläufige und multidimensionale abschließende Fähigkeitsschätzungen nutzen, werden mit einem Algorithmus von multiplen sequentiellen eindimensionalen Tests verglichen. Die eingeschränkten MAT-Algorithmen realisieren dabei eine dimensionsspezifische sequentielle Vorgabe von Items und haben den Anspruch mögliche Übertragungseffekte, die bei der Bearbeitung von Items aus wechselnden Dimensionen auftreten könnten, zu minimieren. Dazu ergibt sich die folgende Fragestellung:

> Forschungsfrage 1.3: Sind die eingeschränkten MAT-Varianten ähnlich präzise, effizient und zuverlässig wie MAT ohne dimensionsspezifische sequentielle Itemvorgabe?

Um bei der Beantwortung der Fragestellungen Erkenntnisse zu gewinnen, die generalisierbar sind, wird die Korrelation zwischen den Dimensionen variiert. Die Verteilung der simulierten Personenparameter soll einer multivariaten Normalverteilung folgen. Die zu simulierenden Itemparameter werden ebenfalls aus einer Normalverteilung gezogen. Zum Vergleich der unterschiedlichen Versuchsbedingungen werden Evaluationskriterien zur Einschätzung der Messpräzision und -effizienz sowie zur Verlässlichkeit der geschätzten Personenparameter berechnet. Die Bearbeitung der Fragestellungen sollte zu direkten Ableitungen hinsichtlich notwendiger Arbeitsschritte der Itempoolentwicklung im Projekt CavE-ICT führen, da die Eigenschaften des später zur Verfügung stehenden Itempools das Ausmaß der Vorteile des adaptiven Testens maßgeblich bestimmen (Frey & Ehmke, 2007).

5.2 Studie II – Erprobung verschiedener CAT-Algorithmen unter Nutzung der CavE-ICT-Feldtestdaten

Im Zuge der Itemkalibrierung im Rahmen des CavE-ICT-Feldtests wurden 70 ICT-Items unter Nutzung eines dichotomen Rasch-Modells skaliert und evaluiert. Davon wurden 64 Items anschließend für den finalen ICT-Itempool selektiert. Neben einem eindimensionalen kam auch ein multidimensionales IRT-Modell zum Einsatz. Dabei zeigte sich, dass aus methodisch-psychometrischer Sicht zur Kalibrierung der Itemparameter das eindimensionale IRT-Modell zu bevorzugen ist. Vor diesem Hintergrund wurde im Projekt CavE-ICT die Entscheidung getroffen, zur Kalibrierung der Itemparameter ein eindimensionales Rasch-Modell zu nutzen. In diesem Sinne werden beispielsweise die spezifische Fähigkeit, in ICT-Umgebungen auf Informationen zugreifen zu können oder verfügbare Informationen bewerten zu können, als unterschiedliche kognitive Prozesse und als Facetten des ICT-Skills-Konstrukts verstanden, nicht aber als

psychometrisch trennbare Dimensionen. Um trotzdem auch differenziert ICT-Personenfähigkeiten für die verschiedenen kognitiven Prozesse rückmelden zu können, wurden multidimensionale Fähigkeitsschätzungen mit fixierten eindimensional geschätzten Itemschwierigkeiten durchgeführt (Kapitel 4). Von einem eher inhaltlich-theoretischen Standpunkt aus gesehen, sind die unterschiedlichen Fähigkeitsausprägungen im Hinblick auf die verschiedenen kognitiven Prozesse, die bei der Lösung von ICT-bezogenen Aufgaben relevant sein können, nämlich durchaus interessant, wissens- und berichtenswert. Diese differenzierteren Informationen sind diagnostisch wertvoll, da sie beispielsweise auch die Empfehlung oder den Einsatz zielgerichteter Interventionen ermöglichen können. Auch andere Leistungstests, insbesondere im Rahmen von LSAs, erfassen Fähigkeiten, von denen anzunehmen ist, dass sie sich aus spezifischeren Teilfähigkeiten oder Facetten zusammensetzen. Dieselben Daten einmal unter Verwendung eines eindimensionalen Modells und einmal unter Verwendung eines mehrdimensionalen Modells zu analysieren, ist beispielsweise im Rahmen von PISA ein durchaus gängiger Ansatz, um alle interessierenden Fähigkeitsschätzungen zu erhalten (OECD, 2009, 2012, 2014).

In Studie II dieser Arbeit sollen die auf diese Weise durch den CavE-ICT-Feldtest gewonnen Informationen in Simulationsstudien genutzt werden, um verschiedene Spezifikationen von CAT-Algorithmen zu prüfen und zu vergleichen. Dazu werden konkret die geschätzten Schwierigkeiten der 64 selektierten ICT-Items sowie die tatsächlich erreichte Verteilung von Items auf die fünf verschiedenen kognitiven Prozesse, die als Merkmalsfacetten von ICT-Skills verstanden werden, einbezogen. Des Weiteren werden auch die Kennwerte der Verteilung der Personenfähigkeiten sowie die gefundenen latenten Korrelationen zwischen den Merkmalsfacetten verwendet. Sogenannte Hybrid-Simulationen (Nydick & Weiss, 2009; Thompson & Weiss, 2011; Weiss & Guyer, 2012), die mit realen Daten – soweit diese vorliegen – und ergänzend mit generierten

Personenfähigkeiten und/oder Antwortmustern arbeiten, ermöglichen eine fundierte Einschätzung darüber, wie sich adaptive Testalgorithmen in Zukunft bei der Administration an echten Personen verhalten werden. Im Rahmen der Entwicklung von CATs werden diese Simulationsstudien daher als unerlässlich angesehen (Schritt 4 nach Thompson & Weiss, 2011). In diesem Sinne sollten, durch den Einbezug von Ergebnissen aus dem CavE-ICT-Feldtest, die Resultate dieser zweiten Simulationsstudie belastbarer sein und konkrete Erkenntnisse zu Einsatzmöglichkeiten von CAT bei der Erfassung von ICT-Skills erlauben.

Die Simulation soll zunächst wieder Fragen zur erforderlichen Länge des adaptiven Tests adressieren. Während in der Studie I die geeignete Testlänge auf Basis von Vorüberlegungen approximiert wurde, soll nun in Studie II das Ergebnis mithilfe der ermittelten Itemeigenschaften und der geschätzten ICT-Fähigkeitsverteilung überprüft werden. Zusätzlich werden Testzeiten aus den bekannten mittleren Itembearbeitungszeiten geschätzt, welche als zusätzliches Effizienzkriterium zur Abschätzung des Zugewinns an Messpräzision bei einer Verlängerung des Tests genutzt werden. Es ergibt sich:

> Forschungsfrage 2.1: Wie viele Items sollten im späteren adaptiven ICT-Test vorgelegt werden, um eine präzise, effiziente und zuverlässige Messung von ICT-Skills zu ermöglichen?

Auf Basis der CavE-ICT-Feldtestdaten soll auch die Effizienz verschiedener Testalgorithmen, die die angenommene multidimensionale Struktur des ICT-Skills-Konstrukts in unterschiedlicher Weise berücksichtigen, noch einmal spezifischer untersucht werden. Neben dem bereits in Studie I genutzten MAT-Algorithmus mit multidimensionaler vorläufiger und abschließender Personenparameterschätzung werden auch wieder ausgewählte eingeschränkte MAT-Algorithmen mit eindimensionaler dimensionsspezifischer vorläufiger und

multidimensionaler abschließender Fähigkeitsschätzung genutzt. Basierend auf den Ergebnissen von Studie I sollten in Studie II allerdings nur noch relevante Algorithmen aufgenommen werden. Verglichen werden diese wiederum mit einem Algorithmus von multiplen sequentiellen eindimensionalen Tests, woraus sich folgende Forschungsfrage ergibt:

> Forschungsfrage 2.2: Sind die eingeschränkten MAT-Varianten auch in einer Simulation mit Echtdaten ähnlich präzise, effizient und zuverlässig wie MAT ohne dimensionsspezifische sequentielle Itemvorgabe?

Ergänzend zu dieser Fragestellung soll weiterhin ein Vergleich zwischen dem MAT-Algorithmus und einem CAT mit genereller eindimensionaler vorläufiger, aber multidimensionaler abschließender Fähigkeitsschätzung vorgenommen werden. Bei beiden Algorithmen ist keine sequentielle Vorgabe der Items pro ICT-Facettenebene vorgesehen. Der Fokus liegt bei dieser Fragestellung vielmehr auf dem Vergleich eines eindimensionalen mit einem multidimensionalen Algorithmus zur Itemauswahl im adaptiven Test. Dies erscheint insofern interessant, als dass der entwickelte ICT-Itempool eindimensional kalibriert wurde, Fähigkeitsschätzungen zukünftig aber mehrdimensional auf Facettenlevel erfolgen sollen. Es stellt sich folgende Frage:

> Forschungsfrage 2.3: Ist ein CAT mit unidimensionaler Itemauswahl ähnlich präzise, effizient und zuverlässig wie MAT, wenn bei beiden Algorithmen die abschließende Fähigkeitsschätzung der fünf Merkmalsfacetten multidimensional erfolgt?

Um die unterschiedlichen Versuchsbedingungen zu vergleichen, werden zur Einschätzung der Messpräzision und -effizienz sowie zur Verlässlichkeit der geschätzten Personenparameter die bereits in Studie I genutzten Evaluationskriterien auch in Studie II herangezogen. Zusätzlich sollen in Studie II die Vorgabehäufigkeit der Items (engl. Exposure-Rates, vgl. Kapitel 3.1.4) und die Einhaltung gesetzter inhaltlicher

Nebenbedingungen (engl. Content-Constraints, vgl. Kapitel 3.1.4) zur Evaluation der verschiedenen Algorithmen herangezogen werden. Zum einen, um zu prüfen, ob bestimmte Test-Algorithmen in besonderem Maße zu einer unerwünschten ungleichmäßigen Vorgabehäufigkeit von Items führen. Zum anderen könnten sich auf diese Weise Items identifizieren lassen, die übermäßig oft vorgegeben werden. Im vorliegenden Fall erscheinen aufgrund des eher kleinen Itempools Methoden zur Kontrolle der Itemvorgabehäufigkeit wenig sinnvoll, da sich nur bei hinreichend großen Itempools die Verteilung der Vorgabehäufigkeit von Items in gewünschtem Maße steuern lässt (Frey, 2012). Nichtsdestotrotz kann die Analyse der Exposure Rates zumindest mögliche Probleme aufdecken. Zum einen können Testalgorithmen dahingehend geprüft werden, ob sie den zur Verfügung stehenden Itempool möglichst gut ausnutzen, zum anderen könnten Items identifiziert werden die besonders häufig in den Tests administriert werden. Stellt sich heraus, dass einzelne Items häufig ausgewählt werden, könnten Alternativen für diese bei einer Erweiterung des ICT-Itempools möglicherweise gezielt nachentwickelt werden.

5.3 Studie III – Zusammenstellung von Kurztests zur eindimensionalen Erfassung von ICT-Skills unter Nutzung der im Zuge des CavE-ICT-Feldtests geschätzten Itemparameter

Das in Kapitel 4 vorgestellte Projekt CavE-ICT liefert, nach intensiver Itementwicklung sowie einer im Rahmen eines Feldtests erfolgten Itemskalierung und -selektion, einen Pool von 64 Items. In der bereits beschriebenen Simulationsstudie II dieser Arbeit wurden die Ergebnisse des Feldtests zur Exploration der Möglichkeiten adaptiven Testens mit dem ICT-Itempool von 64 Items eingesetzt. Auf Basis der CavE-ICT-Projektergebnisse lässt sich aber zunächst auch ableiten, dass der Itempool bei

linearer Testung eine psychometrisch abgesicherte und theoriekonforme Erfassung von ICT-Skills ermöglicht (Wenzel et al., 2016). Dabei muss für den Test allerdings insgesamt eine Bearbeitungszeit von etwa zwei Stunden veranschlagt werden. Ursächlich für diese eher hohe Testzeit ist der Umstand, dass die verhaltensnah umgesetzten Items relativ komplex sind und viele Interaktionsmöglichkeiten bieten (Goldhammer et al., 2014). Zur Bearbeitung sind mitunter ein intensives Explorieren der Aufgabeninhalte oder das Eruieren verschiedener Lösungsmöglichkeiten und -schritte erforderlich. Die Spannbreite der über Personen gemittelten Bearbeitungszeiten von Items liegt zwischen etwa 40 Sekunden und etwas über vier Minuten; der Mittelwert der Bearbeitungszeit über alle Items liegt bei einer Minute und 45 Sekunden mit einer Standardabweichung von 41 Sekunden. Dies zeigt, dass die ICT-Items im Hinblick auf die Bearbeitungszeiten recht heterogen sind. Um ICT-Skills auf ökonomische Weise zu messen, liegt es daher nahe, einen Kurztest zusammenzustellen, der durch die gezielte Auswahl geeigneter Items weniger Zeit bei zugleich hoher Messgenauigkeit und Zuverlässigkeit realisiert. Der aus dem bestehenden ICT-Itempool abgeleitete Kurztest sollte sowohl auf Populationsebene, im Sinne einschlägiger LSAs, als auch in der Gruppen- und Individualdiagnostik einsetzbar sein und daher Anforderungen aus beiden Bereichen gerecht werden. Zielsetzung der Testzusammenstellung im Rahmen von Studie III ist es daher, ein möglichst kurzes, aber dennoch genaues und reliables Instrument zur Messung von ICT-Skills bereitzustellen.

Der Itemauswahlprozess für die zusammenzustellenden ICT-Kurztests erfolgt multikriterial nach verschiedenen Gesichtspunkten. Das durch die Itemauswahl abgedeckte Schwierigkeitsspektrum sollte weitestgehend auch der Fähigkeitsverteilung der Zielpopulation entsprechen. Die Zielpopulation für den Testeinsatz sind 15-jährige Schülerinnen und Schüler. Da die Stichprobe des CavE-ICT-Feldtests aus dieser Population gezogen wurde, kann die aus diesen Daten geschätzte ICT-Fähigkeitsverteilung zur Orientierung bei der Itemselektion herangezogen werden.

Es werden Kurztests zur linearen Erfassung von ICT-Skills mit Testlängen von 10, 15, 20 und 25 Items zusammengestellt, wobei die längeren Tests auf den kürzeren aufbauen, also deren Items enthalten. Die mittleren Bearbeitungszeiten der ausgewählten ICT-Items sollten nach Möglichkeit gering sein, um die benötigte Testzeit insgesamt kurz zu halten. Außerdem sollte die theoretische Konzeption von ICT-Skills durch die Kurztests möglichst umfassend abgebildet werden, um auch für die jeweilig ausgewählte Menge an Items valide Testwertinterpretationen zu ermöglichen. Dazu soll unter anderem in den Kurztests jeweils die gleiche Anzahl von Items für jeden der kognitiven Prozesse, die für die Lösung ICT-spezifischer Aufgaben relevant sein können, administriert werden. Des Weiteren sollte durch die Items eine möglichst große Spannbreite gängiger computerbasierter Anwendungen (Applikationen) abgedeckt werden. Schließlich sollen die für die Kurztests selektierten ICT-Items gut zwischen fähigen und weniger fähigen Personen diskriminieren, wobei hier die Trennschärfe im Sinne der klassischen Testtheorie gemeint ist die als Korrelation von Itemscore und Gesamttestscore erfasst wird und einen Wert von mindesten .30 erreichen sollte (Bortz & Döring, 2006).

Die Testzusammenstellung im Rahmen dieser Studie, bei der aus dem bestehenden ICT-Itempool unter Berücksichtigung der soeben beschriebenen Kriterien Items für Kurztests selektiert werden, kann als ein Problem der eingeschränkten kombinatorischen Optimierung beschrieben werden (van der Linden, 1998b, 2005). In diesem Sinne wird neben manueller Testzusammenstellung auch ein Ansatz der automatisierten computerbasierten Testzusammenstellung verfolgt, bei der lineare Programmierung zum Einsatz kommt. Auf diese Weise werden insgesamt acht ICT-Kurztests zusammengestellt, die anschließend über Simulationsstudien auf die zu erwartende Reliabilität und Messgenauigkeit hin geprüft werden. Dabei werden auch mögliche fehlende Werte in den Simulationen berücksichtigt, um die erzielten Ergebnisse beziehungsweise die aus den Ergebnisse abgeleiteten Erkenntnisse belastbarer zu gestalten. Dazu wird der beobachtete

Anteil fehlender Werte in den im Rahmen des CavE-ICT-Feldtests erhobenen Daten genutzt.

Daraus ergeben sich für Studie III die folgenden Forschungsfragen:

Forschungsfrage 3.1: Unterschieden sich die manuell und die automatisiert zusammengestellten ICT-Kurztests hinsichtlich Messgenauigkeit, -effizienz und Zuverlässigkeit voneinander?

Forschungsfrage 3.2: Wie lang sollte der lineare ICT-Kurztest sein (Testlänge und Testzeit) um eine befriedigend hohe Messgenauigkeit, -effizienz und Zuverlässigkeit der ICT-Fähigkeitsschätzungen zu erzielen?

Zur Einschätzung der Messpräzision und der Zuverlässigkeit der geschätzten Personenparameter, werden die bereits in Studie I und II genutzten Evaluationskriterien auch in Studie III herangezogen. Darüber hinaus wird die Güte der erzielten Personenparameterschätzungen in allen Bereichen des Fähigkeitsspektrums genauer analysiert, um differenzierte Erkenntnisse zur möglichen Nutzung eines ICT-Kurztests im Rahmen von Gruppen- und Individualdiagnostik abzuleiten.

6 Studie I – Spezifikation von Itempoolgröße und Testlänge eines computerisierten adaptiven Tests zur multidimensionalen Messung von ICT-Skills

Im Folgenden werden Methode, Ergebnisse und Diskussion der ersten Studie dieser Arbeit entsprechend der in Kapitel 5.1 vorgestellten Problemstellung und der abgeleiteten Forschungsfragen dargestellt. Zusammengefasst sollte im Rahmen dieser Studie untersucht werden, welchen Einfluss unterschiedlich große Itempools (Forschungsfrage 1.1), die Testlänge (Forschungsfrage 1.2) sowie verschiedene Formen des mehrdimensionalen Testens (Forschungsfrage 1.3) auf die Messpräzision und -effizienz sowie die Zuverlässigkeit eines potentiellen computerisierten adaptiven ICT-Skills Tests haben.

6.1 Methode

Die in Kapitel 5.1 dargestellten Forschungsfragen wurden mithilfe von Monte-Carlo-Simulationen bearbeitet. Bei der Zusammenstellung unterschiedlicher Bedingungen der verschiedenen Simulationen sollten neben Anforderungen, entsprechend der theoretischen Konzeption von ICT-Skills (Kapitel 4), auch mögliche weitere Rahmenbedingungen berücksichtigt werden. Ziel war es bereits vor Beginn der Item- und Testerstellung, Hinweise über die notwendige Größe und Beschaffenheit des zu erarbeitenden Itempools zu generieren. Im Folgenden werden das unter Berücksichtigung dieser Aspekte erstellte Design und die Prozedur der Simulationsstudie beschrieben. Des Weiteren werden die zur Evaluation der verschiedenen Testalgorithmen herangezogenen Kriterien vorgestellt.

6.1.1 Design

Der Versuchsplan der Studie umfasste vier unabhängige Variablen. Für alle Bedingungen wurden $D = 5$ latente Merkmalsdimensionen entsprechend der in der Konstruktbeschreibung von ICT-Skills (Kapitel 4.2.1) dargestellten fünf kognitiven Prozesse, die zur Bearbeitung von ICT-Aufgaben relevant sind, unterschieden. Dabei wurden verschieden hohe *Korrelationen zwischen den Dimensionen* angenommen – da eher starke Zusammenhänge der einzelnen kognitiven Prozesse zu erwarten sind – und in Simulationsbedingungen umgesetzt ($\rho = .50$ und $\rho = .70$). Die *Größe des zur Verfügung stehenden Itempools* wird in der Studie zwischen 50, 75, 100 und 125 Items variiert. Da Einfluss auf die Itementwicklung genommen werden kann, wird von einer idealerweise erreichbaren Gleichverteilung der Items auf die jeweiligen Dimensionen beziehungsweise auf die kognitiven Prozesse, die bei der Bearbeitung von ICT-Aufgaben relevant sind, ausgegangen. Des Weiteren wird die *Testlänge* zwischen 10, 20 und 40 Items variiert. Die vorliegende Simulationsstudie soll Auskunft darüber geben, ob beziehungsweise bis zu welchem Grad der Zugewinn an Messpräzision eine Verlängerung des Tests und damit der Testzeit rechtfertigt. Schließlich wurden fünf verschiedene *Testalgorithmen* verglichen, wobei für vier der Algorithmen abschließende multidimensionale Personenparameterschätzungen für die fünf Merkmalsdimensionen implementiert wurden: multidimensionale adaptive Itemauswahl (Multidimensional Adaptive Testing, MAT) und drei unterschiedliche Varianten sequentieller dimensionsspezifischer adaptiver Itemauswahl (Constrained Multidimensional Adaptive Testing, C-MAT I-III). Zusätzlich wurde ein Algorithmus bestehend aus fünf eindimensionalen Tests für die fünf postulierten Merkmalsdimensionen (Sequential Unidimensional Adaptive Testing, S-UAT) erprobt, wobei die abschließenden Personenparameter für die einzelnen Merkmalsdimensionen jeweils unidimensional geschätzt wurden. Daraus ergab sich ein

vollständiger Versuchsplan mit $3 \times 2 \times 4 \times 5 = 120$ Bedingungen der in Tabelle 6.1 zusammenfassend dargestellt wird.

Tabelle 6.1

Studiendesign zum Vergleich verschiedener Testalgorithmen bei unterschiedlichen Bedingungen hinsichtlich der Testlänge, der Größe des Itempools und der Höhe der Korrelation zwischen Dimensionen

Test-länge	**Korrelation**	**Itempool-größe**	**Finale Fähigkeitsschätzung**				
			Multidimensional				**Unidimensional**
			MAT	C-MAT I	C-MAT II	C-MAT III	S-UAT
10	.50	50					
		75					
		100					
		125					
	.70	50					
		75					
		100					
		125					
20	.50	50					
		75					
		100					
		125					
	.70	50					
		75					
		100					
		125					
40	.50	50					
		75					
		100					
		125					
	.70	50					
		75					
		100					
		125					

Anmerkungen. MAT = Multidimensional Adaptive Testing; C-MAT I-III = Constrained Multidimensional Adaptive Testing I-III; S-UAT = Sequential Unidimensional Adaptive Testing.

6.1.2 Prozedur

Im Folgenden wird die Generierung von Personen- und Itemparametern wie auch der Antwortmatrizen dargestellt. Anschließend wird die Testprozedur eingehender beschrieben. Die Datengenerierung sowie die Simulation der Testprozedur wurden mithilfe der Software SAS (SAS Institute, 2013) durchgeführt.

Datengenerierung. Im ersten Schritt wurden für 5000 Personen Fähigkeitsparameter für fünf Merkmalsdimensionen generiert. Die Verteilung der Personenparameter folgt einer multivariaten Normalverteilung mit $\theta \sim MVN(\mu, \Phi)$, wobei $\mu = (0,0,0,0,0)$ und

$$\Phi = \begin{pmatrix} 1 & \varrho & \varrho & \varrho & \varrho \\ \varrho & 1 & \varrho & \varrho & \varrho \\ \varrho & \varrho & 1 & \varrho & \varrho \\ \varrho & \varrho & \varrho & 1 & \varrho \\ \varrho & \varrho & \varrho & \varrho & 1 \end{pmatrix}. \tag{6.1}$$

Entsprechend der zuvor genannten unterschiedlichen Korrelationsbedingungen ($\varrho = .50$ und $\varrho = .70$) erfolgte die Ziehung von Personenparametern für beide Annahmen. Im nächsten Schritt wurden Itemparameter generiert. Es wurden Pools mit 50, 75, 100 und 125 Items unter Annahme eines fünfdimensionalen Raschmodells in 50 Replikationen erzeugt, wobei jedes Item nur auf exakt eine der fünf Merkmalsdimensionen lädt (Between-Item-Multidimensionalität, Kapitel 2.1.2). Für die Dimensionen war die Anzahl der auf sie ladenden Items jeweils gleich (10, 15, 20 und 25 Items je Dimension). Für jedes Item wurde entsprechend des Modells ein dimensionsspezifischer Schwierigkeitsparameter aus einer eher breit gewählten Normalverteilung mit $b_p \sim N(0,2)$ generiert. Dies entspricht der Zielsetzung der späteren Itementwicklung in dem Sinne, dass auch an den Rändern der Verteilung der Personenfähigkeiten genügend Items für die Gewährleistung einer guten Differenzierbarkeit von Personen vorliegen sollten. Im letzten Schritt wurden unter Verwendung der für die beiden Korrelationsbedingungen erzeugten Personenparameter dichotome Antworten auf die generierten Items für alle Replikationen erzeugt.

Testprozedur. Die Itemauswahl in der *MAT-Bedingung* erfolgte adaptiv nach dem Bayesianischen Ansatz von Segall (1996). Dabei wurde das erste Item für jede Person

zufällig aus dem gesamten zur Verfügung stehenden Itempool gezogen. Danach wurden, basierend auf den Itemparametern, der vorläufigen Fähigkeitsschätzungen einer Person und der latenten Merkmalsverteilung $\mathbf{\Phi}$, Items adaptiv ausgewählt. Dabei wurde jeweils das Item für den Test selektiert, welches den bereits in Formel 3.4 eingeführten Ausdruck

$$\left|\mathbf{I}\left(\theta,\hat{\theta}_T\right)+\mathbf{I}\left(\theta,u_{i*}\right)+\mathbf{\Phi}^{-1}\right| \tag{6.2}$$

maximierte. Dabei entspricht $\mathbf{I}\left(\theta,\hat{\theta}_T\right)$ der Informationsmatrix der T bereits administrierten Items und $\mathbf{I}\left(\theta,u_{i*}\right)$ der Informationsmatrix eines noch verfügbaren Items $i*$. Die Reihenfolge der vorgelegten Items ist demnach nicht dimensionsspezifisch, sondern folgt ausschließlich dem Kriterium maximaler Iteminformation. Um dennoch den Anteil der vorgelegten Items pro Dimension im Test gleich zu verteilen, wurde – im Sinne des Content Managements (Kapitel 3.1.4) – der Multidimensional-Maximum-Priority-Index (MMPI; Frey et al., 2011; siehe Kapitel 3.2.2) implementiert. Im Gegensatz zur MAT-Bedingung wurden in den C-MAT-Bedingungen die Items sequentiell für jede Dimension nacheinander adaptiv ausgewählt. Variiert wurden die Informationen, die jeweils zur Auswahl des ersten Items einer neuen Dimension einbezogen wurden, wobei das erste Item des Gesamttests in allen C-MAT-Bedingungen zufällig aus den für die erste Dimension zur Verfügung stehenden Items ausgewählt wurde. Danach erfolgte die Itemauswahl adaptiv auf Basis der Itemparameter und der vorläufigen unidimensionalen Personenparameterschätzung für die spezifische Dimension. Sobald eine Anzahl von $1/D \cdot$ Testlänge Items und demnach im vorliegenden Fall 20% der im Test insgesamt zu administrierenden Items vorgelegt wurden, wurde zur nächsten Dimension übergegangen. In der *C-MAT I-Bedingung* erfolgte die Auswahl des ersten Testitems sowie der jeweils ersten Items für die weiteren Dimensionen zufällig und die weitere dimensionsspezifische Itemauswahl entsprechend

der Itemparameter und vorläufigen Personenparameterschätzung allein auf dieser Dimension. Somit handelte es sich in dieser Bedingung um sequentiell vorgegebene unidimensionale adaptive Tests, bei der weder die Korrelations-Matrix $\mathbf{\Phi}$ der multidimensionalen latenten Merkmalsverteilung noch die Antworten auf Items vorheriger Dimensionen in die vorläufige Fähigkeitsschätzung und Itemauswahl einbezogen wurden. Dagegen wurden in der *C-MAT II-Bedingung* beim Dimensionswechsel die Personenparameterschätzung aus der vorherigen Dimension und die angenommene Korrelation zwischen den Dimensionen für die Auswahl des ersten Items der jeweiligen neu zu administrierenden Dimension genutzt. Die weitere Selektion von Items für die jeweilige Dimension erfolgte wiederum eindimensional adaptiv auf Basis der Itemparameter und der vorläufigen dimensionsspezifischen Fähigkeitsschätzung. In der *C-MAT III-Bedingung* wurden beim Wechsel der Dimension die Personenparameterschätzung aller bisher administrierten Dimensionen und die multidimensionale latente Merkmalsverteilung $\mathbf{\Phi}$ für die Auswahl des ersten Items der jeweiligen neuen Dimension genutzt. Auch in dieser Bedingung erfolgte die weitere Itemselektion auf Basis der gegebenen Itemparameter und der vorläufigen dimensionsspezifisch geschätzten Personenparameter jeweils unidimensional adaptiv. Die Personenparameter wurden am Ende der Testadministration für die Bedingungen MAT und C-MAT I-III basierend auf den erzeugten Antworten und der Korrelations-Matrix $\mathbf{\Phi}$ der latenten Merkmalsverteilung geschätzt. In diesen Bedingungen wurde demnach die Multidimensionalität des Konstrukts im Sinne einer abschließenden multidimensionalen Fähigkeitsschätzung der Personen einbezogen. Die S-*UAT-Bedingung* soll als Vergleichsbedingung ohne multidimensionale Fähigkeitsschätzungen dienen. Die Itemauswahl erfolgte auf die gleiche Weise wie in C-MAT-Bedingung I bereits beschrieben, allerdings werden die abschließenden Personenparameter für jede Dimension nicht unter Einbezug der Korrelation der Merkmalsdimensionen geschätzt.

Als Personenparameter wurden MAP-Schätzwerte (siehe Kapitel 2) bestimmt. Tabelle 6.2 gibt eine Zusammenschau und bietet einen Vergleich über die verschiedenen Testalgorithmen.

Tabelle 6.2
Zusammenschau und Vergleich der verschiedenen Testalgorithmen

Testalgorithmus	MAT	C-MAT I	C-MAT II	C-MAT III	S-UAT
Reihenfolge der Dimensionen	Gemischt	Nacheinander von $D=1$ bis $D=5$			
Auswahl des ersten Items einer neuen Dimension ($D=2$ bis $D=5$)	-	Zufällig	Auf Basis der θ-Schätzung der zuletzt administrierten Dimension und der Korrelation zwischen den Dimensionen	Auf Basis der θ-Schätzungen aller zuvor administrierten Dimensionen und der Korrelation zwischen den Dimensionen	Zufällig
Vorläufige θ-Schätzung zur Itemauswahl	Multi-dimensional	Ein-dimensional			
Abschließende θ-Schätzung	Multi-dimensional				Ein-dimensional

Anmerkungen. MAT = Multidimensional Adaptive Testing; C-MAT I-III = Constrained Multidimensional Adaptive Testing I-III; S-UAT = Sequential Unidimensional Adaptive Testing.

Die Testadministration wurde entsprechend der jeweiligen Versuchsbedingung nach Vorlage von 10, 20 bzw. 40 Items beendet. Für jede im Studiendesign vorgesehene Versuchsbedingung erfolgte die Simulation der vorgestellten Testalgorithmen in 50 Replikationen mit 5000 Personen.

6.1.3 Evaluationskriterien

Zum Vergleich der Testbedingungen wurden vier verschiedene Evaluationskriterien berechnet, die im Sinne des Studiendesigns als abhängige Variablen verstanden werden können. Berechnet wurde der systematische Messfehler (engl. *Bias*) der

Fähigkeitsschätzungen als Mittelwert der Differenz zwischen dem geschätzten und wahren Personenparameter jeder Dimension:

$$Bias_d = \frac{\sum_{j=1}^{N}\left(\hat{\theta}_{jd} - \theta_{jd}\right)}{N} \quad . \tag{6.3}$$

Dabei entspricht $\hat{\theta}_{jd}$ der geschätzten Fähigkeit und θ_{jd} der wahren Fähigkeit für Person j auf der Merkmalsdimension d. Berichtet wird der mittlere Bias über alle Replikationen für jede der fünf Merkmalsdimensionen ebenso wie über die Dimensionen gemittelt. Als Maß für die Messpräzision wird die mittlere quadratische Abweichung, (englisch *Mean Squared Error,* MSE) ebenfalls als Durchschnittswert über die Replikationen berichtet. Der MSE einer Dimension d ergibt sich aus der durchschnittlichen quadrierten Abweichung der geschätzten und wahren Personenparameter dieser Dimension:

$$MSE_d = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_{jd} - \theta_{jd}\right)^2 . \tag{6.4}$$

Gemittelt über die Dimensionen ergibt sich:

$$MSE = \frac{1}{K}\sum_{d=1}^{D} MSE_d . \tag{6.5}$$

Zum Vergleich der verschiedenen Testalgorithmen wurde auch ein Maß der relativen Messeffizienz (engl. *Relative Efficiency*, RE; vgl. (Kröhne et al., 2014; La Torre & Patz, 2005) berechnet. Zur Untersuchung des Effekts der abschließenden multidimensionalen Personenparameterschätzung wurde der MSE der S-UAT Bedingung als Vergleichsbasis gewählt und den MSEs der Bedingungen MAT und C-MAT I-III (als alternative Testalgorithmen) gegenübergestellt:

$$RE = \frac{MSE_{\text{S-UAT}}}{MSE_{\text{Alternativer Algorithmus}}} . \tag{6.6}$$

Ein RE größer 1 impliziert, dass der alternative Algorithmus effizienter ist als der Testalgorithmus der Vergleichsbedingung (S-UAT). Die *Reliabilität* (REL) der Personenparameterschätzungen gibt einen Hinweis auf die praktische Relevanz erhöhter Messeffizienz (Kröhne et al., 2014) und wurde als quadrierte Korrelation geschätzter und wahrer Personenparameter für jede Dimension d ermittelt:

$$REL_d = r^2_{\hat{\theta}_{jd},\theta_{jd}} . \tag{6.7}$$

Auch bei der Berechnung der Reliabilität wurde für jede Dimension über die Replikationen gemittelt und ein Mittelwert über die Dimensionen erstellt. Bei diesen Berechnungen wurde zunächst eine Fisher-Z-Transformation (Fisher, 1915) der quadrierten Korrelationen vorgenommen und nach der Mittelwertsberechnung wieder rücktransformiert. Zur Einschätzung der Reliabilität wird sich an den im Rahmen von PISA-Studien berichteten Werten orientiert (vgl. Mikolajetz, 2017). In den Jahren 2009 und 2012 (OECD, 2009, 2012, 2014) lagen die, auf Basis von eindimensionalen Skalierungen der einzelnen Fähigkeitsbereiche (Mathematics, Reading, Science), errechneten Reliabilitäten für die Hauptdomänen in einem Bereich von .85 bis .86 und für die jeweiligen Nebendomänen zwischen .54 und .57. Bei Verwendung Multidimensionaler Schätzungen und unter Einbezug eines Hintergrundmodells konnten auch für die Nebendomänen höhere Reliabilitätswerte erzielt werden. In den neueren Technischen Berichten der PISA-Studien 2015 (OECD, 2018) und 2018 werden nur noch diese Reliabilitäten angegeben, die Werte zwischen .85 und .93 annehmen. Nichts desto trotz wird im Rahmen der vorliegenden Studie eine Reliabilität von .55 als Mindestschwelle für das Berichten von Testwerten in LSAs angenommen.

6.2 Ergebnisse

Im Folgenden werden für die einzelnen Evaluationskriterien über Dimensionen gemittelte Werte angegeben. Die Werte für die einzelnen Dimensionen sind jeweils den entsprechenden Anhängen zu entnehmen. Die Bildung eines Mittelwerts über Dimensionen ist im Rahmen dieser Studie zulässig, da die durch die Datensimulation gesetzten Voraussetzungen (Verteilungsannahmen und Höhe der Korrelationen zwischen Dimensionen) für alle Dimensionen gleichgesetzt sind. Daher können sich die Ergebnisse je Dimension nicht wesentlich voneinander unterschieden. Um einen besseren Überblick über die Ergebnisse zu erlangen, können daher über Dimensionen gemittelte Werte berichtet und zur Einschätzung der Güte der jeweiligen Fähigkeitsschätzung der einzelnen Dimensionen herangezogen werden.

6.2.1 Systematischer Messfehler – Bias

Für alle Testbedingungen wurden kleine Werte für den Bias gefunden. Die über Replikationen und Dimensionen gemittelten Werte des Bias sowie der zugehörige Standardfehler für die verschiedenen Versuchsbedingungen und Testalgorithmen können Tabelle 6.3 entnommen werden. Zudem sind im Anhang B.1 ausdifferenzierte Tabellen mit einer Aufschlüsselung des Bias nach Dimension angefügt. Die Werte des Testalgorithmus mit unidimensionalen vorläufigen und abschließenden Fähigkeitsschätzern (S-UAT) waren insgesamt etwas größer als für die multidimensionalen Tests, wobei für die verschiedenen Testalgorithmen mit multidimensionalen Fähigkeitsschätzungen (MAT, C-MAT I-III) keine systematischen Unterschiede zwischen den Algorithmen beobachtet werden konnten. Bei steigender Testlänge wurden die Werte des Bias über alle Testalgorithmen hinweg etwas weniger von Null verschieden, wobei weder die Höhe der Korrelation zwischen den latenten Merkmalsdimensionen noch die Größe des Itempools einen substantiellen Einfluss zu

haben schienen. Insgesamt lässt sich festhalten, dass die Fähigkeitsschätzungen im Mittel unverzerrt (engl. unbiased) waren.

Tabelle 6.3

Über die Replikationen und Dimensionen gemittelter Bias und Standardfehler (SE)

Test-länge	Kor-relation	Item-pool-größe	Testalgorithmen:				
			MAT Bias (*SE*)	C-MAT I Bias (*SE*)	C-MAT II Bias (*SE*)	C-MAT III Bias (*SE*)	S-UAT Bias (*SE*)
10	.50	50	.001 (.007)	.003 (.008)	.002 (.007)	.002 (.008)	.003 (.009)
		75	.001 (.006)	.002 (.007)	.002 (.007)	.003 (.006)	.003 (.008)
		100	.002 (.007)	.002 (.008)	.001 (.007)	.002 (.007)	.002 (.007)
		125	.000 (.007)	.001 (.007)	.001 (.006)	.001 (.007)	.002 (.007)
	.70	50	-.003 (.005)	-.002 (.007)	-.003 (.007)	-.002 (.008)	-.003 (.008)
		75	-.001 (.006)	-.002 (.007)	-.001 (.006)	-.001 (.007)	-.004 (.007)
		100	-.003 (.006)	-.002 (.006)	-.003 (.007)	-.002 (.006)	-.004 (.007)
		125	-.003 (.005)	-.002 (.007)	-.003 (.007)	-.003 (.006)	-.003 (.007)
20	.50	50	.001 (.007)	.001 (.008)	.001 (.008)	.001 (.008)	.002 (.009)
		75	.001 (.006)	.002 (.006)	.001 (.007)	.003 (.006)	.003 (.007)
		100	.001 (.007)	.001 (.007)	-.000 (.007)	.000 (.007)	.002 (.007)
		125	.000 (.007)	.001 (.006)	.002 (.006)	-.001 (.007)	.002 (.007)
	.70	50	-.001 (.006)	-.001 (.007)	-.001 (.006)	-.002 (.006)	-.003 (.009)
		75	-.001 (.005)	-.002 (.006)	-.000 (.006)	-.001 (.006)	-.003 (.007)
		100	-.003 (.006)	-.002 (.006)	-.002 (.005)	-.002 (.006)	-.003 (.007)
		125	-.002 (.005)	-.002 (.006)	-.002 (.006)	-.002 (.006)	-.003 (.007)
40	.50	50	.001 (.011)	.001 (.011)	.001 (.011)	.001 (.011)	.001 (.012)
		75	.001 (.007)	.001 (.007)	.001 (.007)	.002 (.007)	.001 (.008)
		100	-.000 (.007)	.000 (.007)	.000 (.007)	.000 (.007)	.000 (.008)
		125	.000 (.006)	.000 (.006)	.000 (.006)	-.000 (.006)	.000 (.007)
	.70	50	-.001 (.008)	-.001 (.008)	-.001 (.008)	-.001 (.008)	-.002 (.012)
		75	.000 (.006)	-.000 (.006)	-.000 (.006)	-.000 (.006)	-.002 (.012)
		100	-.001 (.006)	-.002 (.006)	-.001 (.006)	-.001 (.006)	-.003 (.008)
		125	-.001 (.005)	-.001 (.006)	-.002 (.006)	-.002 (.005)	-.002 (.007)

Anmerkungen. MAT = Multidimensional Adaptive Testing; C-MAT I-III = Constrained Multidimensional Adaptive Testing I-III; S-UAT = Sequential Unidimensional Adaptive Testing.

6.2.2 Mittlere quadrierte Abweichung – Mean Squared Error (MSE)

In Tabelle 6.4 sind die MSEs in den verschiedenen Versuchsbedingungen mit Standardfehlern abgetragen. Dem Anhang B.2 kann zusätzlich eine genauere Aufschlüsselung der MSEs nach Dimensionen entnommen werden.

Tabelle 6.4

Über die Replikationen und Dimensionen gemittelter Mean Squared Error (MSE) und Standardfehler (SE)

Test-länge	Kor-relation	Item-pool-größe	Testalgorithmen:				
			MAT *MSE* (*SE*)	C-MAT I *MSE* (*SE*)	C-MAT II *MSE* (*SE*)	C-MAT III *MSE* (*SE*)	S-UAT *MSE* (*SE*)
10	.50	50	.599 (.012)	.620 (.011)	.603 (.009)	.603 (.009)	.725 (.013)
		75	.597 (.010)	.619 (.010)	.601 (.009)	.601 (.009)	.724 (.012)
		100	.597 (.009)	.617 (.009)	.600 (.008)	.602 (.009)	.723 (.011)
		125	.597 (.008)	.618 (.010)	.599 (.009)	.601 (.009)	.723 (.011)
	.70	50	.504 (.010)	.528 (.010)	.509 (.009)	.510 (.009)	.718 (.013)
		75	.503 (.008)	.527 (.009)	.507 (.008)	.509 (.008)	.717 (.012)
		100	.502 (.009)	.527 (.010)	.506 (.009)	.509 (.009)	.717 (.011)
		125	.501 (.008)	.525 (.009)	.506 (.008)	.508 (.008)	.716 (.010)
20	.50	50	.454 (.018)	.465 (.009)	.457 (.009)	.458 (.009)	.557 (.011)
		75	.450 (.011)	.462 (.008)	.454 (.008)	.454 (.008)	.554 (.011)
		100	.448 (.009)	.461 (.008)	.451 (.007)	.453 (.007)	.552 (.009)
		125	.447 (.008)	.460 (.008)	.451 (.008)	.452 (.007)	.551 (.010)
	.70	50	.370 (.012)	.383 (.008)	.375 (.007)	.376 (.007)	.553 (.013)
		75	.367 (.008)	.381 (.007)	.372 (.007)	.374 (.006)	.549 (.011)
		100	.365 (.007)	.379 (.007)	.370 (.006)	.372 (.006)	.547 (.010)
		125	.365 (.006)	.378 (.007)	.370 (.007)	.372 (.006)	.547 (.009)
40	.50	50	.341 (.019)	.344 (.013)	.343 (.013)	.343 (.013)	.407 (.017)
		75	.320 (.016)	.325 (.008)	.322 (.008)	.322 (.008)	.383 (.011)
		100	.313 (.011)	.319 (.007)	.316 (.006)	.316 (.007)	.376 (.008)
		125	.310 (.008)	.316 (.006)	.313 (.006)	.313 (.006)	.372 (.007)
	.70	50	.280 (.013)	.283 (.009)	.282 (.009)	.282 (.009)	.405 (.018)
		75	.263 (.011)	.268 (.006)	.266 (.006)	.266 (.006)	.380 (.011)
		100	.257 (.008)	.263 (.005)	.261 (.005)	.262 (.005)	.373 (.008)
		125	.255 (.006)	.262 (.005)	.258 (.004)	.259 (.005)	.370 (.008)

Anmerkungen. MAT = Multidimensional Adaptive Testing; C-MAT I-III = Constrained Multidimensional Adaptive Testing I-III; S-UAT = Sequential Unidimensional Adaptive Testing.

Es zeigt sich, dass die erreichten MSEs für alle vier multidimensionalen Testalgorithmen (MAT und C-MAT I-III) sehr ähnlich waren, wobei von diesen der C-MAT I, welcher am wenigsten Informationen zur Itemauswahl im Testverlauf nutzt, auch am wenigsten gut abschnitt. Mit steigender Testlänge verringerten sich die MSEs für alle Testalgorithmen deutlich. Für den MAT-Algorithmus mit multidimensionalen vorläufigen und abschließenden Fähigkeitsschätzungen wurden die jeweils niedrigsten Werte des MSE beobachtet. Hingegen fallen die MSEs für den S-UAT mit unidimensionalen vorläufigen und abschließenden Fähigkeitsschätzern über alle Testlängen hinweg am höchsten aus. Die höhere Korrelation der einzelnen latenten

Merkmalsdimensionen wirkte sich erwartungsgemäß nur auf die MSEs der multidimensionalen Testalgorithmen verringernd aus. Die Anzahl von zur Verfügung stehenden Items schien einen eher geringen Einfluss auf den MSE zu haben, der erst bei längeren Tests zunahm.

6.2.3 Relative Messeffizienz – Relative Efficiancy (RE)

Ein Vergleich der relativen Messeffizienz der verschiedenen multidimensionalen adaptiven Testalgorithmen in Bezug zum S-UAT wird in Abbildung 6.1 für die verschiedenen Testlängen sowie nach Höhe der Korrelation zwischen den latenten Merkmalsdimensionen und der Größe des Itempools dargestellt. Im Anhang B.3 sind die Werte der relativen Messeffizienz auch nach Dimensionen inklusive der Standardfehler einsehbar.

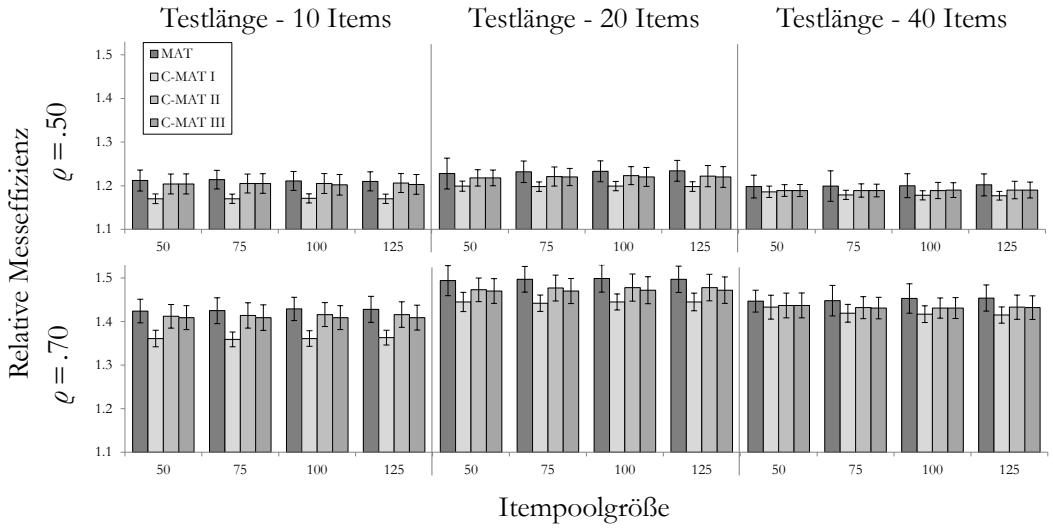

Abbildung 6.1 Relative Messeffizienz in den verschiedenen Versuchsbedingungen für die unterschiedlichen Testalgorithmen (MAT = Multidimensional Adaptive Testing; C-MAT I-III = Constrained Multidimensional Adaptive Testing I-III).

Die relative Messeffizienz zeigte in allen Versuchsbedingungen deutliche Vorteile der multidimensionalen Tests zum eindimensionalen Test (alle Werte liegen deutlich über 1). Dabei erzielte der MAT in allen Versuchsbedingungen die höchsten Werte, die bei einer Korrelation der latenten Merkmalsdimensionen von .50 zwischen 1.198 und 1.234 und bei einer Korrelation von .70 zwischen 1.424 und 1.499 lagen. Die Testalgorithmen mit eindimensionaler vorläufiger und multidimensionaler abschließender Fähigkeitsschätzung wiesen in allen Versuchsbedingungen eine niedrigere relative Messeffizienz auf als der MAT. C-MAT II und III unterschieden sich, mit Werten zwischen 1.189 und 1.223 bei zu .50 korrelierten Merkmalsdimensionen und 1.409 und 1.478 bei Korrelationen von .70, nicht substantiell hinsichtlich ihrer Relativen Effizienz. Der C-MAT I-Algorithmus erzielte die niedrigsten Werte (1.170 – 1.199 bei $\varrho = .50$ und 1.361 – 1.445 bei $\varrho = .70$). Es zeigte sich, dass die Vorteile der multidimensionalen Testalgorithmen im Vergleich zum S-UAT-Algorithmus sich besonders dann entfalten, wenn der Test eine mittlere Länge von 20 Items hat und die latenten Merkmalsdimensionen eher hoch korreliert sind.

6.2.4 Reliabilität

In Abbildung 6.2 werden die über Replikationen und Dimensionen gemittelten Reliabilitäten der verschiedenen Testalgorithmen grafisch dargestellt. Anhang B.4 kann eine detaillierte Aufschlüsselung der Reliabilitäten nach Dimensionen sowie mit zugehörenden Standardfehlern entnommen werden.

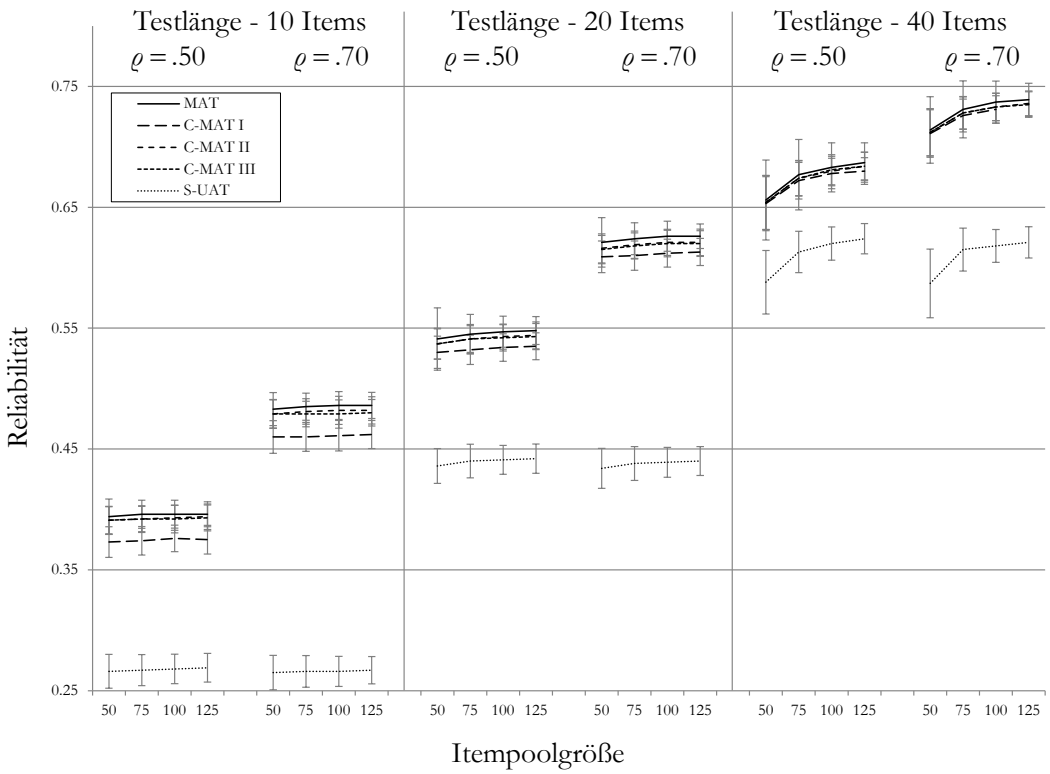

Abbildung 6.2 Reliabilitäten für die unterschiedlichen Testalgorithmen je nach Versuchsbedingung (MAT = Multidimensional Adaptive Testing; C-MAT I-III = Constrained Multidimensional Adaptive Testing I-III; S-UAT = Sequential Unidimensional Adaptive Testing).

Die Reliabilität lag für den eindimensionalen Testalgorithmus (S-UAT) in allen Versuchsbedingungen deutlich unter den multidimensionalen Tests (MAT und C-MAT I-III). In der S-UAT-Bedingung konnten Reliabilitäten von maximal .269 für 10 Items, .440 für 20 Items und .624 für 40 Items Testlänge erreicht werden. Mit Blick auf die erreichten Reliabilitäten der multidimensionalen Testalgorithmen zeigte sich, dass MAT, C-MAT II und C-MAT III ähnliche Werte erzielten. Diese lagen bei hoher Korrelation der latenten Dimensionen und einer Testlänge von 10 Items zwischen .479 und .486, bei 20 Items zwischen .615 und .626 und bei 40 Items zwischen .712 und .739. Mit zunehmender Testlänge stiegen die Reliabilitäten erwartungsgemäß an und wurden zudem

zwischen den einzelnen Testalgorithmen ähnlicher. Während die Werte für die C-MAT I-Bedingungen mit maximal .462 bei einer Testlänge von 10 Items noch deutlich unter den anderen multidimensionalen Testalgorithmen lagen, waren sie bei einer Testlänge von 40 Items mit maximal .732 nicht mehr deutlich niedriger. Die Höhe der Korrelation der verschiedenen latenten Merkmalsdimensionen hatte einen positiven Effekt auf die Reliabilität der multidimensionalen Testalgorithmen (MAT und C-MAT I-III). Die Größe des zur Verfügung stehenden Itempools schien erst bei hinreichend langen Tests von größerer Bedeutung zu sein.

6.3 Diskussion

Im Rahmen der ersten Studie der vorliegenden Arbeit wurde, unter Einsatz von Simulationen, untersucht, wie der zu entwickelnde ICT-Itempool aussehen und wie lang ein möglicher adaptiver Test sein sollte, der zudem die angenommene Mehrdimensionalität von ICT-Skills berücksichtigt. Neben der Messpräzision und Messeffizienz wurde auch die Reliabilität verschiedener Testalgorithmen analysiert. Die erzielten Ergebnisse zeigen in Bezug auf die in Kapitel 5.1 aufgestellten Forschungsfragen 1.1 und 1.2, dass die Größe des für den adaptiven Test zur Verfügung stehenden Itempools erst mit steigender Anzahl vorzulegender Items ins Gewicht fällt. Dies lässt sich vor allem mit Blick auf die Werte der MSEs und der Reliabilitäten ableiten. Die erzielten Verringerungen der MSEs für die verschiedenen adaptiven Testalgorithmen durch eine Vergrößerung des Itempools von 50 auf bis zu 125 Items sind erst bei einer Testlänge von 40 vorzulegenden Items als bedeutsam anzusehen. Ähnliches gilt für die Reliabilitäten, wobei hier nur zwischen den Bedingungen mit einem Itempool von lediglich 50 Items zu denen mit größeren Itempools bedeutsame Verbesserungen zu beobachten sind. Die Höhe der Korrelation der latenten Merkmalsdimensionen wirkt sich zudem bei Testalgorithmen, die diese Mehrdimensionalität in die Itemauswahl und/oder

Fähigkeitsschätzungen einbeziehen, positiv auf MSE und Reliabilität aus. Um die anvisierte Reliabilitätsmindestschwelle von .50 für das Berichten von Testwerten im Rahmen von LSAs zu erreichen, sollte der spätere adaptive ICT-Test bei Einbezug der Mehrdimensionalität von ICT-Skills mindestens 20 Items lang sein.

Zur Untersuchung der dritten Forschungsfrage dieser Studie (1.3) wurden vier multidimensionale adaptive Testalgorithmen, von denen drei die Items geordnet für eine Dimension nach der anderen vorgeben, mit einem sequentiellen unidimensionalen Algorithmus verglichen. Dazu wurde unter anderem die relative Messeffizienz als Verhältnis des MESs des unidimensionalen Algorithmus zu den multidimensionalen Algorithmen berechnet. Dabei zeigt sich, dass sich die Vorteile der MAT- und C-MAT-Algorithmen bei einer mittleren Testlänge von 20 Items und einer eher hohen Korrelation der latenten Merkmalsdimensionen am deutlichsten entfalten. Zwar zeigt sich der MAT-Algorithmus in allen Versuchsbedingungen erwartungsgemäß mit der besten Messpräzision, Messeffektivität und Reliabilität, allerdings liegen vor allem die C-MAT-Bedingungen II und III sehr nah an den erzielten Werten des MAT-Algorithmus. Somit bieten diese Algorithmen mit sequentieller Vorgabe von Items für eine Dimension nach der anderen tatsächlich eine mögliche Alternative zum MAT-Algorithmus, sofern die Wichtigkeit der sequentiellen Vorgabe als ausreichend hoch erachtet wird, um die geringen Einbußen in Messpräzision, Messeffektivität und Reliabilität auszugleichen. Es konnte differenziert gezeigt werden, dass die C-MAT-Varianten besser funktionieren, wenn mehr Informationen zum Zusammenhang der Merkmalsdimensionen zur Auswahl der jeweiligen dimensionsspezifischen Startitems herangezogen werden. So weist der C-MAT I-Algorithmus, in dem solche Informationen gar nicht genutzt werden, die unter den multidimensionalen Algorithmen schlechtesten Ergebnisse in Bezug auf alle angewendeten Evaluationskriterien auf. Substantielle Unterschiede zwischen den Varianten C-MAT II und III lassen sich hingegen nur in sehr wenigen Bedingungen bei

kleinerer Testlänge für Bias und MSE beobachten. Generell lässt sich aber konstatieren, dass beide Bedingungen, gegeben der simulierten Datenstruktur, ähnlich gut funktionieren. In Bezug auf die umgesetzten Testalgorithmen lässt sich damit festhalten, dass – sofern eine sequentielle Vorgabe von Items geordnet nach Dimensionen gewünscht ist – die hier vorgestellten Alternativen zum MAT genutzt werden können. Dabei sollte aber zur Auswahl der jeweiligen dimensionsspezifischen Startitems auf Informationen im Hinblick auf die Korrelation der Merkmalsdimensionen und der auf zuvor administrierten Dimensionen geschätzten Personenfähigkeiten zurückgegriffen werden. Wird dies berücksichtigt, können mit MAT vergleichbar präzise und reliable Ergebnisse erzielt werden. Dabei gilt es jedoch jeweils abzuwägen, ob die geringen Einbußen bei MSE und Reliabilität zugunsten der sequentiellen dimensionsspezifischen Itemvorgabe verkraftbar sind.

Ziel der vorgestellten ersten Simulationsstudie dieser Arbeit war es, Anforderungen für die Konstruktion des ICT-Itempools abzuleiten. Aus den erzielten Ergebnissen lässt sich schließen, dass nach der Kalibrierung der ICT-Items im Rahmen des CavE-ICT-Feldtests zwischen 50 und 75 zum Raschmodell gut passende Items zur Verfügung stehen sollten, um in einem adaptiven ICT-Skills-Test eingesetzt zu werden. Dieser Itempool sollte zudem das Schwierigkeitsspektrum möglichst breit abbilden, um für die individuelle Merkmalsausprägung jedes Testteilnehmers hinreichend viele trennscharfe Items mit adäquater Schwierigkeit auswählen zu können. Um des Weiteren das zu messende Konstrukt angemessen zu repräsentieren, sollte bereits bei der Itemkonstruktion darauf abgezielt werden, eine annähernde Gleichverteilung von Items zu den fünf kognitiven Prozessen zu realisieren. Um die Zielgröße von 50 bis 75 Items zu erreichen, sollten allerdings etwa 100 Items konstruiert und im Feldtest eingesetzt werden, da davon auszugehen ist, dass ein Teil dieser Items beispielsweise auf Grund schlechter

Modellpassung oder unzureichender Diskrimination nicht für den ICT-Itempool selektiert werden können.

In Bezug auf die Länge des späteren adaptiven ICT-Skills-Tests kann festgehalten werden, dass längere Tests mit einer Verringerung der Messfehler und Verbesserung der Reliabilität einhergehen. Für die im Rahmen dieser Studie untersuchten multidimensionalen Tests zeigt sich allerdings, dass schon eine Testlänge von 20 Items, also nur vier vorgelegten Items je Dimension, zu Reliabilitäten führt, die im Rahmen von LSAs für das Berichten von Testwerten als ausreichend angesehen werden können. Zudem überwiegen bei dieser Testlänge die Vorteile der multidimensionalen Testalgorithmen zum unidimensionalen Test, beobachtet anhand der relativen Messeffizienz, am deutlichsten. Mit zunehmender Testlänge nehmen diese deutlichen Vorteile des multidimensionalen Tests wieder etwas ab. Die tatsächliche Relevanz der Testlänge muss allerdings im Anschluss an den CavE-ICT-Feldtest erneut beurteilt werden, da erst dann belastbare Informationen zur notwendigen Bearbeitungszeit der ICT-Items vorliegen. Die Länge des späteren adaptiven ICT-Skills-Tests sollte dann in Abwägung von zu erreichender Messpräzision und Reliabilität mit verfügbarer oder notwendiger Testzeit getroffen werden. Zudem gilt es im Rahmen des CavE-ICT-Feldtests zu prüfen, wie hoch die Korrelationen der latenten Merkmalsdimensionen tatsächlich ausfallen, da auch diese Einfluss auf die erforderliche Testlänge haben. Die vorliegende Simulationsstudie gibt zunächst Auskunft darüber, dass eine Länge von lediglich 10 Items unzureichend und von 40 Items eventuell mehr als notwendig ist. Im Anschluss an den CavE-ICT-Feldtest werden auf Basis der im Rahmen dieser Studie erworbenen Erkenntnisse und der dann verfügbaren Daten, unter anderem zu Item- und Personenparametern, sogenannte Hybrid-Simulationen (Nydick & Weiss, 2009; Thompson & Weiss, 2011; Weiss & Guyer, 2012) durchgeführt. In dieser zweiten Studie wird der C-MAT II-Algorithmus nicht mehr zum Einsatz kommen, da sich die

unter Einsatz des C-MAT II und III erzielten Ergebnisse kaum unterschieden. Zudem repräsentiert der C-MAT III-Algorithmus stärker die Nutzung von zur Verfügung stehenden Informationen und stellt daher einen stärkeren Kontrast zum C-MAT I-Algorithmus her. Die untersuchten möglichen Testlängen sollen dann 10, 20, 30 und 40 Items betragen, um potentielle Zugewinne durch eine Testverlängerung differenzierter detektieren und abbilden zu können. Des Weiteren werden zusätzliche Evaluationskriterien herangezogen, um den Nutzen von möglichen Testspezifikationen genauer zu untersuchen.

7 Studie II – Erprobung verschiedener CAT-Algorithmen unter Nutzung der CavE-ICT-Feldtestdaten

In diesem Kapitel wird auf die Bearbeitung der in Kapitel 5.2 hergeleiteten Fragestellungen abgezielt. Hierfür werden im Folgenden Methode, Ergebnisse und Diskussion zur zweiten Studie dieser Arbeit vorgestellt. Im Zuge dieser Studie wird anhand der CavE-ICT-Feldtestdaten exploriert, wie der erarbeitete ICT-Itempool im Rahmen des adaptiven Testens genutzt werden kann. Dazu wird untersucht, welchen Einfluss die Testlänge (Forschungsfrage 2.1) und verschiedene Formen des mehrdimensionalen Testens (Forschungsfrage 2.2 und 2.3) auf die Messpräzision und -effizienz sowie die Zuverlässigkeit eines potentiellen computerisierten adaptiven ICT-Skills Test haben. Daneben werden auch Einschränkungen an die Testzusammenstellung näher betrachtet. Zum einen wird das Funktionieren der in einigen Algorithmen zur Anwendung gebrachten Content Management Methoden untersucht und zum anderen Exposure Rates analysiert.

7.1 Methode

Um die in Kapitel 5.2 dargestellten Forschungsfragen zu untersuchen, wurden auch im Rahmen dieser Studie Monte-Carlo-Simulationen durchgeführt. Da aus dem CavE-ICT-Feldtest ein IRT-kalibrierter Itempool und ICT-Fähigkeitsschätzungen von 766 Schülerinnen und Schülern vorliegen, sollten diese Daten genutzt werden, um die Performanz unterschiedlicher CAT-Algorithmen mit unterschiedlichen Testlängen zu evaluieren. Dieses Vorgehen kann als eine Art von Hybrid-Simulationen (Nydick & Weiss, 2009; Thompson & Weiss, 2011; Weiss & Guyer, 2012) verstanden werden. Im vorliegenden Fall wurden die geschätzten Itemparameter aus dem CavE-ICT-Feldtest verwendet und Personenfähigkeiten unter Nutzung der im Feldtest beobachteten

Fähigkeitsverteilung sowie Antwortmatrizen generiert. Wie in Kapitel 4.3 und 5.2 beschrieben, wurden die Itemparameter unter Nutzung eines eindimensionalen Raschmodells geschätzt. Die Schätzung der Personenparameter für die kognitiven Prozesse, die bei der Bearbeitung ICT-bezogener Aufgaben relevant sind, erfolgte mithilfe fixierter Itemparameter unter Nutzung eines fünfdimensionalen Raschmodells. Die kognitiven Prozesse werden inhaltlich als Ebenen einer Merkmalsfacette von ICT-Skills und nicht als psychometrisch trennbare Dimensionen angesehen (vgl. Kapitel 4.2.1). Im Folgenden wird dennoch der Begriff der *Merkmalsdimensionen* genutzt, da die technische Umsetzung der Itemauswahl und Fähigkeitsschätzung in den verschiedenen angewendeten CAT-Algorithmen einer Auffassung von Merkmalsdimensionen folgt. Im Folgenden werden zunächst das Design und die Prozedur der Simulationsstudie beschrieben. Des Weiteren werden die zur Evaluation der verschiedenen Testalgorithmen herangezogenen Kriterien vorgestellt.

7.1.1 Design

Der Versuchsplan dieser Studie hatte zwei unabhängige Variablen. Zum einen wurde, wie im Rahmen von Studie I, die *Testlänge* variiert, um Auskunft darüber zu erhalten, wie stark die Messpräzision durch die Verlängerung des Tests – und damit der Testzeit – zunimmt und sich längere Tests für eine genauere Testung rechtfertigen lassen. Allerdings wurden nun Testlängen von 10, 20, 30 und 40 Items untersucht, um noch differenziertere Aussagen über die benötigte vorzugebende Itemanzahl in einem adaptiven ICT-Skills-Test treffen zu können. Zum anderen wurden verschiedene adaptive *Testalgorithmen* verglichen. Für alle Bedingungen wurden $D = 5$ latente Merkmalsdimensionen angenommen. Diese Dimensionen entsprechend den fünf bei der Bearbeitung von ICT-Aufgaben relevanten kognitiven Prozessen entsprechend der Rahmenkonzeption von ICT-Skills (Kapitel 4.2.1). Äquivalent zur Studie I wurde ein Algorithmus mit multidimensionaler adaptiver Itemauswahl (Multidimensional Adaptive

Testing, MAT) angewendet. Zusätzlich wurde ein Algorithmus eingesetzt, der auf eindimensionale adaptive Itemauswahl zurückgreift (Combined Unidimensional Multidimensional Adaptive Testing, CU-MAT). Schließlich wurden zwei der in Studie I genutzten Varianten sequentieller dimensionsspezifischer adaptiver Itemauswahl (Constrained Multidimensional Adaptive Testing I und III, C-MAT I und III) auch in dieser Studie angewendet. Für alle vier Algorithmen wurden abschließende multidimensionale Personenparameterschätzungen implementiert. Zusätzlich wurde, ebenfalls wie in Studie I, ein Algorithmus bestehend aus fünf eindimensionalen Tests für die fünf postulierten Merkmalsfacetten (Sequential Unidimensional Adaptive Testing, S-UAT) umgesetzt, bei dem die abschließenden Personenparameter für die einzelnen Merkmalsdimensionen jeweils unidimensional geschätzt wurden. Es ergibt sich ein vollständiger Versuchsplan mit $4 \times 5 = 20$ Bedingungen, der in Tabelle 7.1 abgebildet wird.

Tabelle 7.1
Studiendesign zum Vergleich verschiedener Testalgorithmen bei unterschiedlichen Testlängen

Testlänge	**Finale Fähigkeitsschätzung**				
	Multidimensional				**Unidimensional**
	MAT	CU-MAT	C-MAT I	C-MAT III	S-UAT
10					
20					
30					
40					

Anmerkungen. MAT = Multidimensional Adaptive Testing; CU-MAT = Combined Unidimensional Multidimensional Adaptive Testing; C-MAT I/III = Constrained Multidimensional Adaptive Testing I/III; S-UAT = Sequential Unidimensional Adaptive Testing.

7.1.2 Prozedur

Die Datengenerierung wie auch die Simulation der Testprozedur wurden mithilfe der Software SAS (SAS Institute, 2013) durchgeführt.

Datengenerierung. Zunächst wurden für 1000 Personen Fähigkeitsparameter auf fünf Merkmalsdimensionen (entsprechend der fünf kognitiven Prozesse) generiert, wobei die empirische Verteilung aus dem CavE-ICT-Feldtest herangezogen wurde. Die Verteilung der Personenparameter folgt einer multivarianten Verteilung mit $\theta \sim MV(\mu, \boldsymbol{\Sigma})$, wobei $\mu = (0,0,0,0,0)$ und die aus den Feldtestdaten hervorgehende folgende Varianz-Kovarianz-Matrix (vgl. Tabelle 4.3 in Kapitel 4) genutzt werden:

$$\boldsymbol{\Sigma} = \begin{pmatrix} .42 & .34 & .30 & .19 & .24 \\ .34 & .53 & .41 & .26 & .33 \\ .30 & .41 & .61 & .25 & .33 \\ .19 & .26 & .25 & .36 & .22 \\ .24 & .33 & .33 & .22 & .53 \end{pmatrix}. \tag{7.1}$$

Anschließend wurden unter Verwendung der generierten Personenparameter Antworten der 1000 Personen auf die 64 ICT-Items in 100 Replikationen erzeugt.

Testprozedur. Wie in Studie I erfolgte die Itemauswahl in der *MAT-Bedingung* adaptiv nach dem Bayesianischen Ansatz von Segall (1996). Dabei wurde das erste Item für jede Person zufällig aus dem zur Verfügung stehenden ICT-Itempool gezogen. Danach wurden Items adaptiv ausgewählt, basierend auf den ICT-Itemparametern, der vorläufigen Fähigkeitsschätzung einer Person und der aus den ICT-Feldtestdaten berechneten latenten Korrelation der kognitiven Prozesse Zugreifen, Managen, Integrieren, Bewerten und Erzeugen (vgl. Tabelle 4.3 in Kapitel 4):

$$\boldsymbol{\Phi} = \begin{pmatrix} 1 & .72 & .59 & .48 & .51 \\ .72 & 1 & .72 & .59 & .63 \\ .59 & .72 & 1 & .53 & .58 \\ .48 & .59 & .53 & 1 & .50 \\ .51 & .63 & .58 & .50 & 1 \end{pmatrix}. \tag{7.2}$$

Die Reihenfolge der vorgelegten Items folgt ausschließlich dem Kriterium maximaler Iteminformation und ist damit nicht dimensionsspezifisch. Um den Anteil

der vorgelegten Items pro Dimension im Test gleich zu verteilen, wurde wie in Studie I für die MAT-Bedingung der MMPI (Frey et al., 2011; siehe Kapitel 3.2.2) implementiert. Auch in der *CU-MAT-Bedingung* erfolgt die Itemauswahl nach dem Kriterium der maximalen Information nicht dimensionsspezifisch. Allerdings werden die Items basierend auf den ICT-Itemparametern und der vorläufigen Fähigkeitsschätzung eindimensional adaptiv aus dem gesamten ICT-Itempool ausgewählt. Als Constraint Management Methode für eindimensionale CATs wurde der Maximum-Priority-Index (MPI; Y. Cheng & Chang, 2009; siehe Kapitel 3.1.4) für die CU-MAT-Bedingung implementiert. Die Personenparameter wurden am Ende der Testadministration basierend auf den erzeugten Antworten und der Korrelations-Matrix $\mathbf{\Phi}$ der latenten multivariaten Merkmalsverteilung für jede Dimension beziehungsweise Merkmalsfacette geschätzt. Im Gegensatz zur MAT- und CU-MAT-Bedingung wurden in den beiden *C-MAT-Bedingungen* die Items sequentiell für jede Dimension nacheinander adaptiv ausgewählt. Das erste Item des Gesamttests wurde zufällig aus den für die erste Dimension zur Verfügung stehenden ICT-Items ausgewählt, danach erfolgte die Itemauswahl auf Basis der ICT-Itemparameter und der vorläufigen unidimensionalen Personenparameterschätzung für die spezifische Dimension adaptiv. Sobald eine Anzahl von $1/D \cdot$ Testlänge Items vorgelegt wurden, wurde zur nächsten Dimension übergegangen. In der *C-MAT I-Bedingung* erfolgte auch die Auswahl der erstens Items aller weiteren Dimensionen zufällig. In der *C-MAT III-Bedingung* wurde beim Dimensionswechsel die Personenparameterschätzung aller bisher administrierten Dimensionen und die Korrelation zwischen den Dimensionen ($\mathbf{\Phi}$) für die Auswahl des ersten Items der jeweiligen neu zu administrierenden Dimension genutzt. Die weitere Selektion von Items für die jeweilige Dimension erfolgte in beiden C-MAT-Bedingungen auf Basis der ICT-Itemparameter und der vorläufigen dimensionsspezifischen Fähigkeitsschätzung eindimensional adaptiv. Wie in der CU-MAT-Bedingung wurden

die Personenparameter am Ende der Testadministration basierend auf den erzeugten Antworten und der Korrelations-Matrix **Φ** multidimensional geschätzt. Die *S-UAT-Bedingung* dient, wie in Studie I, als Vergleichsbedingung ohne multidimensionale Fähigkeitsschätzungen. Die Auswahl des ersten Testitems sowie der jeweils ersten Items für die weiteren Dimensionen erfolgte zufällig und die weitere dimensionsspezifische Itemauswahl entsprechend der ICT-Itemparameter und vorläufigen Personenparameterschätzung allein auf dieser Dimension. Somit handelte es sich in dieser Bedingung um sequentiell vorgegebene unidimensionale adaptive Tests. Als Personenparameter wurden jeweils MAP-Schätzwerte (Kapitel 2.2) bestimmt. Tabelle 7.2 gibt einen Überblick zu den verschiedenen Testalgorithmen.

Tabelle 7.2

Zusammenschau und Vergleich der verschiedenen in Studie II implementierten Testalgorithmen

Testalgorithmus	MAT	CU-MAT	C-MAT I	C-MAT III	S-UAT
Reihenfolge der Dimensionen	Gemischt		Nacheinander von $D = 1$ bis $D = 5$		
Auswahl des ersten Items einer neuen Dimension ($D = 2$ bis $D = 5$)	-		Zufällig	Auf Basis der θ-Schätzungen aller zuvor administrierten Dimensionen und der Korrelation zwischen den Dimensionen	Zufällig
Vorläufige θ-Schätzung zur Itemauswahl	Multi-dimensional	Ein-dimensional (gesamt)		Ein-dimensional (dimensionsspezifisch)	
Abschließende θ-Schätzung	Multi-dimensional				Ein-dimensional

Anmerkungen. MAT = Multidimensional Adaptive Testing; CU-MAT = Combined Unidimensional Multidimensional Adaptive Testing; C-MAT I & III = Constrained Multidimensional Adaptive Testing I & III; S-UAT = Sequential Unidimensional Adaptive Testing.

Die Testadministration wurde entsprechend der jeweiligen Versuchsbedingung nach Vorlage von 10, 20, 30 und 40 Items beendet. Für jede im Studiendesign vorgesehene Versuchsbedingung erfolgte die Simulation der vorgestellten Testalgorithmen in 100 Replikationen mit jeweils 1000 Personen.

7.1.3 Evaluationskriterien

Zum Vergleich der verschiedenen Testbedingungen wurden verschiedene Evaluationskriterien berechnet. Wie bereits in Studie I (Kapitel 6) wurden *Bias* (siehe Formel 6.3), *MSE* (siehe Formel 6.4), *Relative Messeffizienz* (*RE*; siehe Formel 6.6) und *Reliabilität* (siehe Formel 6.7) der verschiedenen Testalgorithmen für die unterschiedlichen Bedingungen und für jede der fünf Dimensionen beziehungsweise Merkmalsfacetten berechnet und über die 100 Replikationen gemittelt. Wie in Studie I wird RE zur Untersuchung des Effekts der abschließenden multidimensionalen Personenparameterschätzung genutzt. Daher wurde der MSE der S-UAT-Bedingung wiederum als Vergleichsbasis gewählt und den MSEs der Bedingungen MAT, CU-MAT, C-MAT I und C-MAT III (als alternative Testalgorithmen) gegenübergestellt. Zur Beurteilung der Reliabilität gelten ebenfalls, wie in Studie I, auch im Rahmen dieser Studie Werte von .55 als Mindestschwelle für den Bericht von Testwerten im Kontext von LSAs (vgl. Kapitel 6.1.3). Für die Ermittlung der über die Replikationen gemittelten Reliabilitäten wurden Fisher-Z-Transformationen genutzt.

Um zu prüfen, ob die durch die Implementierung des MPI und des MMPI gesetzten Bedingungen in Bezug auf die Vorgabehäufigkeit von Items eingehalten wurden, wurde für die beiden Algorithmen mit diesen Constraint-Management-Methoden (CU-MAT und MAT) der Anteil von korrekt administrierten Items und Tests betrachtet. Zum einen wurde der prozentuale Anteil von Tests, bei denen der Anteil der vorgelegten Items pro Dimension gleichverteilt ist, ermittelt und zum anderen die prozentuale Anzahl von Items, die entsprechend der Nebenbedingungen administriert wurden, angegeben.

Des Weiteren sollte auch die Vorgabehäufigkeit der Items (engl. *Exposure-Rate*) näher betrachtet werden. Dazu wurden die absoluten Item Exposure-Rates über alle Replikationen betrachtet. Berichtet wird zudem die Anzahl nicht-administrierter Items, da nicht vorgelegte Items auf eine nicht optimale Ausnutzung des zur Verfügung stehenden Itempools hinweisen. Vor diesem Hintergrund wurde zudem ein Maß der Testüberlappung (engl. *Test-Overlap-Rate*, TOR; vgl. Chen, Ankenmann & Spray, 2003; Y. Cheng & Chang, 2009) berechnet. Dieser Index verdeutlicht, wie häufig die gleichen Items in Tests zweier unterschiedlicher Personen auftreten. Chen et al. (2003) zeigen, dass TOR aus den Item-Exposure-Rates, der Testlänge und der Anzahl administrierter adaptiver Tests mit fixer Testlänge berechnet werden kann:

$$TOR = \frac{N \cdot \sum_{i=1}^{M} \left(\frac{T_i}{N} \right)^2}{L \cdot (N-1)} - \frac{1}{N-1}. \tag{7.3}$$

Vereinfacht ergibt sich dafür:

$$TOR = \frac{\sum_{i=1}^{M} T_i (T_i - 1)}{L \cdot N (N-1)}. \tag{7.4}$$

T_i gibt die Anzahl von Administrationen eines Items i an, M bezeichnet die Größe des Itempools, L die Testlänge und N die Stichprobengröße. Die TOR wird für jede Replikation berechnet und zur Ergebnisdarstellung über die Replikationen gemittelt. Eine hohe TOR deutet auf ein hohes Maß an Testüberlappung hin und somit auf eine weniger gute Ausnutzung des zur Verfügung stehenden Itempools als bei niedrigerer TOR.

Abschließend sollen die im Mittel zu erwartenden *Testzeiten* für die vier Testlängen unabhängig vom eingesetzten CAT-Algorithmus berichtet werden, um eine Aussage zur einzuplanenden Erhebungszeit beim praktischen Einsatz eines adaptiven ICT-Tests zu ermöglichen.

7.2 Ergebnisse

7.2.1 Systematischer Messfehler – Bias

In Tabelle 7.3 sind die über Replikationen gemittelten Werte des Bias sowie der zugehörige Standardfehler für die fünf Dimensionen nach Testlängen und -algorithmen abgebildet.

Tabelle 7.3

Über die Replikationen gemittelter Bias und Standardfehler (SE) für jede der fünf Merkmalsdimensionen

Testlänge	Dimension	Testalgorithmen: MAT	CU-MAT	C-MAT I	C-MAT III	S-UAT
		Bias (*SE*)	Bias (*SE*)	Bias (*SE*)	Bias (*SE*)	Bias (*SE*)
10	1	- .018 (.013)	- .020 (.013)	- .018 (.012)	- .021 (.012)	- .023 (.011)
	2	- .015 (.013)	- .015 (.014)	- .016 (.013)	- .016 (.014)	- .024 (.015)
	3	- .010 (.013)	- .010 (.015)	- .010 (.014)	- .011 (.013)	- .017 (.014)
	4	.006 (.013)	.002 (.014)	.002 (.013)	- .001 (.014)	- .005 (.014)
	5	- .013 (.014)	- .012 (.015)	- .013 (.013)	- .013 (.014)	- .030 (.014)
20	1	- .014 (.013)	- .014 (.012)	- .014 (.011)	- .015 (.012)	- .019 (.014)
	2	- .010 (.013)	- .009 (.012)	- .010 (.013)	- .010 (.013)	- .020 (.017)
	3	- .006 (.013)	- .006 (.013)	- .005 (.013)	- .005 (.013)	- .016 (.016)
	4	.005 (.013)	.003 (.013)	.004 (.013)	.003 (.012)	- .002 (.015)
	5	- .010 (.014)	- .007 (.013)	- .008 (.014)	- .006 (.014)	- .022 (.016)
30	1	- .011 (.012)	- .011 (.012)	- .011 (.012)	- .012 (.012)	- .022 (.014)
	2	- .007 (.012)	- .008 (.011)	- .008 (.012)	- .008 (.012)	- .014 (.017)
	3	- .004 (.012)	- .003 (.012)	- .002 (.013)	- .003 (.012)	- .013 (.016)
	4	.006 (.012)	.003 (.012)	.004 (.012)	.003 (.012)	- .003 (.015)
	5	- .006 (.013)	- .005 (.013)	- .005 (.012)	- .005 (.012)	- .023 (.014)
40	1	- .010 (.012)	- .008 (.012)	- .008 (.011)	- .008 (.011)	- .016 (.014)
	2	- .004 (.010)	- .006 (.010)	- .006 (.011)	- .006 (.012)	- .011 (.016)
	3	- .002 (.012)	- .000 (.012)	- .000 (.012)	- .000 (.012)	- .005 (.016)
	4	.006 (.013)	.003 (.012)	.003 (.012)	.003 (.013)	.002 (.014)
	5	- .003 (.012)	- .004 (.012)	- .004 (.012)	- .004 (.012)	- .031 (.014)

Anmerkungen. MAT = Multidimensional Adaptive Testing; CU-MAT = Combined Unidimensional Multidimensional Adaptive Testing; C-MAT I & III = Constrained Multidimensional Adaptive Testing I & III; S-UAT = Sequential Unidimensional Adaptive Testing; Dimensionen = Kognitive Prozesse: 1 = Zugreifen, 2 = Managen, 3 = Integrieren, 4 = Bewerten, 5 = Erzeugen.

Mit zunehmender Testlänge werden die Werte des Bias über alle Testalgorithmen hinweg geringer. Der Bias in den S-UAT-Bedingungen liegt für einzelne Dimensionen deutlich höher als in den Testalgorithmen mit multidimensionaler abschließender Fähigkeitsschätzung; in den jeweiligen S-UAT-Bedingungen zeigen sich auch die größten Schwankungen der Bias-Werte zwischen den Dimensionen. Die Werte des Bias fallen bei allen Testalgorithmen und vor allem für kürzere Testlängen (10 und 20 Items) besonders niedrig aus für die vierte Merkmalsdimension „Bewerten". Die erste Dimension „Zugreifen" zeigt bei Testalgorithmen mit multidimensionaler abschließender Fähigkeitsschätzung jeweils höhere Bias-Werte. Ab einer Testlänge von 30 Items können für alle Dimensionen bei Testalgorithmen mit multidimensionaler abschließender Fähigkeitsschätzung unverzerrte Fähigkeitsschätzungen beobachtet werden (beurteilt unter Berücksichtigung des jeweiligen Standardfehlers des Bias).

7.2.2 Mittlere quadrierte Abweichung – Mean Squared Error (MSE)

In Tabelle 7.4 sind die Werte des MSE über die Replikationen gemittelt sowie der Standardfehler des MSE für die fünf Dimensionen und alle Versuchsbedingungen abgetragen. Über alle eingesetzten Algorithmen hinweg fällt der MSE für die kürzeste Testlänge (10 Items) bei Dimension vier „Bewerten" am geringsten aus. Die MSEs für die Dimensionen drei „Integrieren" und fünf „Erzeugen" fallen über alle Testalgorithmen hinweg hingegen am höchsten aus.

Tabelle 7.4

Über die Replikationen gemittelter Mean Squared Error (MSE) und Standardfehler (SE) für jede der fünf Merkmalsdimensionen

Testlänge	Dimension	Testalgorithmen:				
		MAT	CU-MAT	C-MAT I	C-MAT III	S-UAT
		MSE (*SE*)	*MSE* (*SE*)	*MSE* (*SE*)	*MSE* (*SE*)	*MSE* (*SE*)
10	1	.359 (.014)	.355 (.014)	.364 (.017)	.355 (.013)	.434 (.017)
	2	.383 (.015)	.367 (.016)	.380 (.017)	.365 (.013)	.475 (.017)
	3	.439 (.016)	.435 (.017)	.453 (.015)	.433 (.016)	.542 (.016)
	4	.325 (.013)	.328 (.013)	.333 (.014)	.326 (.015)	.388 (.015)
	5	.412 (.015)	.415 (.016)	.429 (.013)	.412 (.015)	.502 (.016)
20	1	.287 (.011)	.287 (.012)	.292 (.011)	.288 (.010)	.383 (.015)
	2	.288 (.011)	.275 (.012)	.282 (.013)	.275 (.010)	.404 (.017)
	3	.335 (.011)	.331 (.012)	.340 (.013)	.331 (.013)	.440 (.017)
	4	.274 (.012)	.279 (.011)	.283 (.012)	.279 (.012)	.354 (.016)
	5	.328 (.014)	.329 (.014)	.335 (.014)	.328 (.013)	.418 (.017)
30	1	.248 (.010)	.248 (.010)	.251 (.010)	.250 (.009)	.337 (.014)
	2	.226 (.009)	.228 (.009)	.232 (.009)	.227 (.009)	.347 (.014)
	3	.284 (.011)	.278 (.010)	.283 (.011)	.278 (.010)	.373 (.015)
	4	.243 (.011)	.248 (.011)	.251 (.010)	.248 (.010)	.316 (.013)
	5	.285 (.013)	.282 (.011)	.287 (.012)	.282 (.013)	.363 (.015)
40	1	.221 (.010)	.224 (.010)	.227 (.010)	.226 (.009)	.305 (.012)
	2	.187 (.007)	.200 (.009)	.203 (.008)	.201 (.007)	.304 (.013)
	3	.255 (.009)	.251 (.010)	.253 (.010)	.252 (.010)	.336 (.015)
	4	.226 (.010)	.228 (.011)	.229 (.010)	.228 (.010)	.287 (.013)
	5	.270 (.012)	.263 (.011)	.264 (.012)	.264 (.011)	.336 (.014)

Anmerkungen. MAT = Multidimensional Adaptive Testing; CU-MAT = Combined Unidimensional Multidimensional Adaptive Testing; C-MAT I & III = Constrained Multidimensional Adaptive Testing I & III; S-UAT = Sequential Unidimensional Adaptive Testing; Dimensionen = Kognitive Prozesse: 1 = Zugreifen, 2 = Managen, 3 = Integrieren, 4 = Bewerten, 5 = Erzeugen.

Mit zunehmender Testlänge zeigen sich aber für Dimension zwei „Managen“ die niedrigsten Werte des MSE. Für diese Dimension stehen auch die meisten Items zur Verfügung. Für die verschiedenen Bedingungen mit multidimensionalen Testalgorithmen können zudem niedrigere MSEs beobachtet werden als für die mit unidimensionalem Testalgorithmus. Mit zunehmender Testlänge sinken den MSEs erwartungsgemäß in allen Bedingungen.

7.2.3 Relative Messeffizienz – Relative Efficiancy (RE)

Der Vergleich der relativen Messeffizienz der verschiedenen MAT-Algorithmen in Bezug zum S-UAT wird in Abbildung 7.1 für die fünf Merkmalsdimensionen und verschiedenen Testlängen dargestellt.

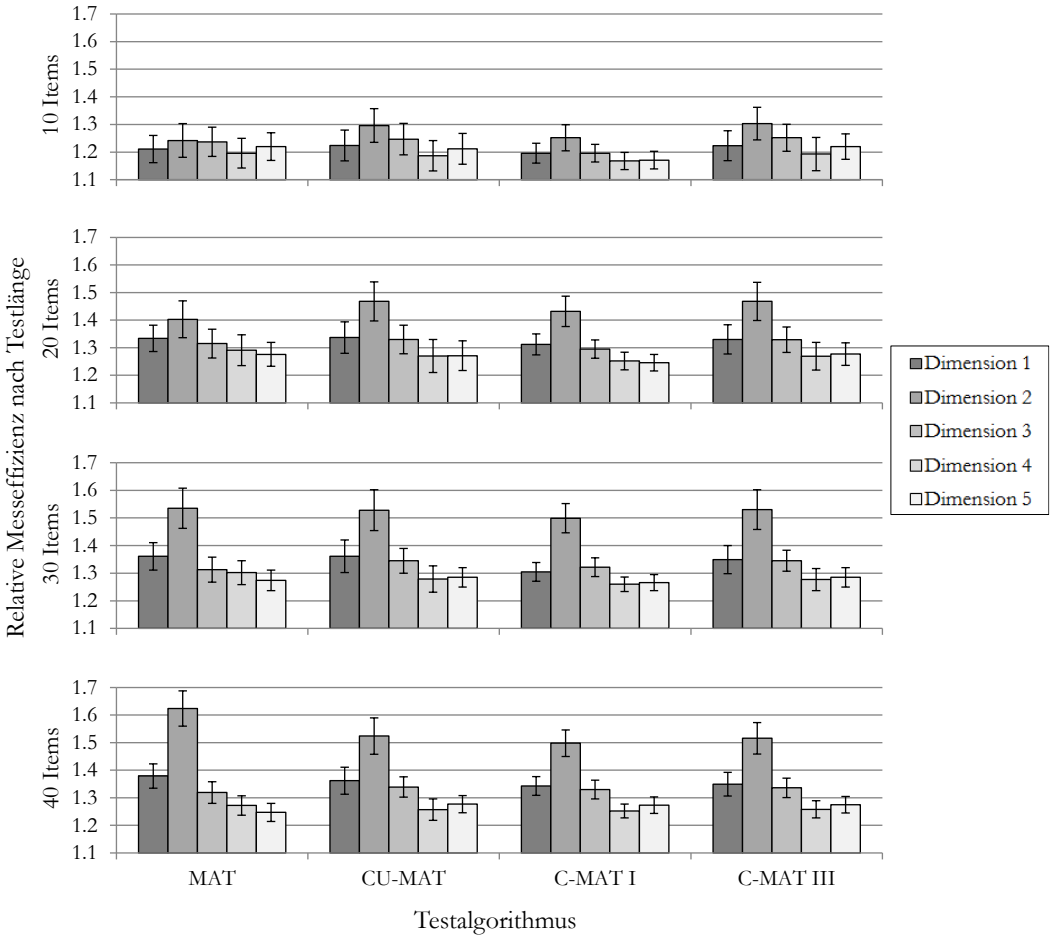

Abbildung 7.1 Relative Messeffizienz der unterschiedlichen Testalgorithmen (MAT = Multidimensional Adaptive Testing; C-MAT I & III = Constrained Multidimensional Adaptive Testing I & III; S-UAT = Sequential Unidimensional Adaptive Testing) bei verschiedenen Testlängen für die fünf Merkmalsdimensionen (Dimensionen = Kognitive Prozesse: 1 = Zugreifen, 2 = Managen, 3 = Integrieren, 4 = Bewerten, 5 = Erzeugen).

Die einzelnen Werte der RE für alle Versuchsbedingungen und jede Dimension können Anhang C.1 entnommen werden. Alle Werte für die RE liegen deutlich über 1, was auf Vorteile der multidimensionalen Testalgorithmen im Vergleich zum S-UAT hindeutet. Die RE steigt für alle Testalgorithmen mit Blick auf alle Dimensionen deutlich von Tests mit 10 Items zu denen mit 20 und mehr Items Testlänge. Dabei zeigen sich für die C-MAT I-Bedingungen die vergleichsweise niedrigsten RE-Werte. Bei einer Testlänge von 30 Items zeigen sich die höchsten RE-Werte, bei einer weiteren Erhöhung der Testlänge steigert sich die RE dann nur noch für einzelne Dimensionen bei der Verwendung des MAT-Algorithmus. Es zeigen sich über alle verwendeten Testalgorithmen hinweg die besten RE-Werte für Dimension zwei „Managen". Für diese Dimension schwanken die Werte auch am stärksten; hier sind die größten Zuwächse an RE bei Verlängerung der Tests von 10 auf 20 und 30 Items zu erkennen.

7.2.4 Reliabilität

In Abbildung 7.2 werden die Reliabilitäten der Fähigkeitsschätzungen der einzelnen Merkmalsdimensionen dargestellt. Die einzelnen Reliabilitäten für alle Versuchsbedingungen und jede Dimension können Anhang C.2 entnommen werden. Die Abbildung zeigt deutlich, dass sich die Reliabilitäten der dimensionsspezifischen Fähigkeitsschätzungen unterscheiden.

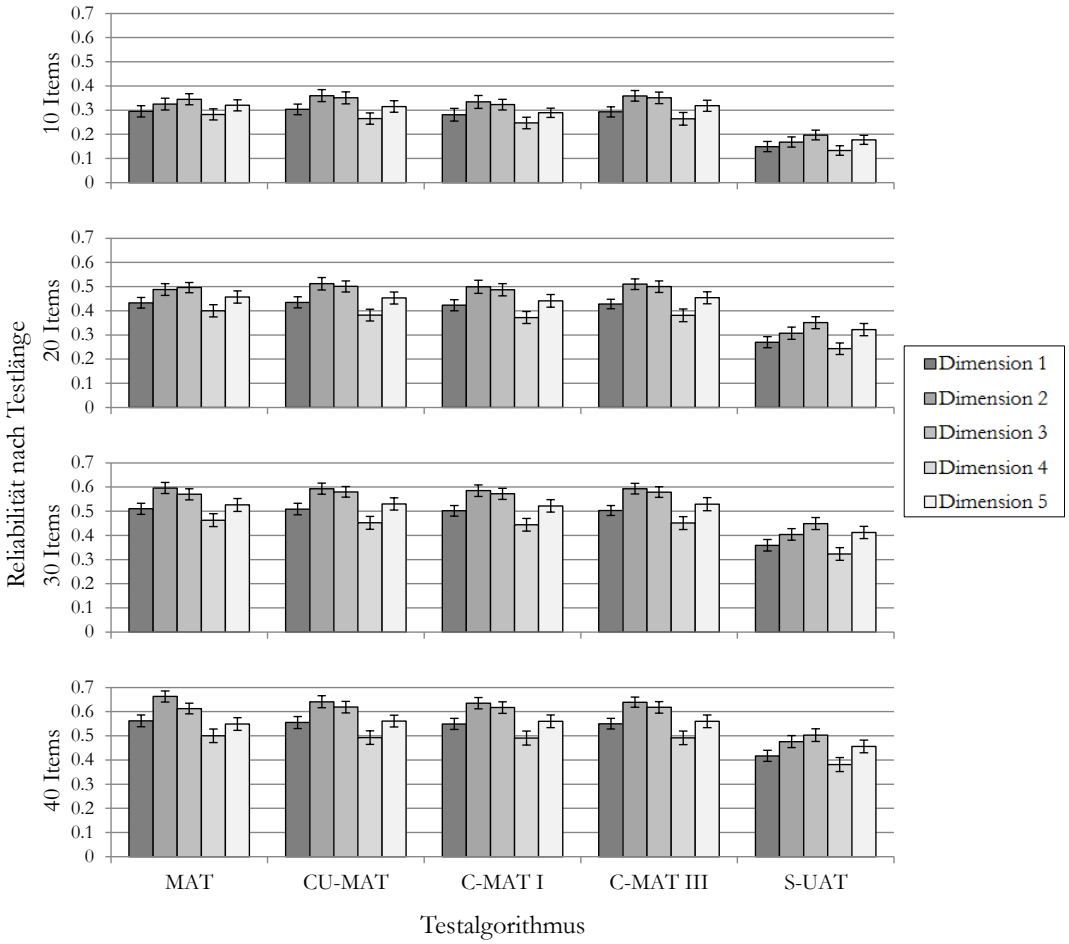

Abbildung 7.2 Reliabilitäten der unterschiedlichen Testalgorithmen (MAT = Multidimensional Adaptive Testing; C-MAT I & III = Constrained Multidimensional Adaptive Testing I & III; S-UAT = Sequential Unidimensional Adaptive Testing) bei verschiedenen Testlängen für die fünf Merkmalsdimensionen (Dimensionen = Kognitive Prozesse: 1 = Zugreifen, 2 = Managen, 3 = Integrieren, 4 = Bewerten, 5 = Erzeugen).

Es zeigen sich über alle verwendeten Testalgorithmen und Testlängen hinweg für Dimension vier „Bewerten" die schlechtesten Reliabilitäten, die bei einer Testlänge von 40 Items zwischen .381 für die S-UAT-Bedingung und .500 für die MAT-Bedingung liegen. Die Dimensionen zwei „Managen" und drei „Integrieren" weisen hingegen bessere Reliabilitäten auf. Für Testalgorithmen mit multidimensionaler vorläufiger und/oder

abschließender Fähigkeitsschätzung liegen die Werte bei einer Testlänge von 30 Items zwischen .585 und .596 für „Managen“ und zwischen .570 und .580 für „Integrieren“. Bei einer Testlänge von 40 Items liegen diese Werte sogar deutlich über .600; für „Managen“ zwischen .635 und .663 und für „Integrieren“ zwischen .613 und .619. Die niedrigsten Reliabilitätswerte konnten für den kognitiven Prozess „Bewerten“ beobachteten werden, diese liegen bei einer Testlänge von 40 ICT-Items bei .500 (MAT), .493 (CU-MAT), .491 (C-MAT I), .492 (C-MAT II) und .381 (S-UAT). In der S-UAT-Bedingung wurden für jede untersuchte Testlänge und geschätzte Dimension die niedrigsten Reliabilitäten erzielt (zwischen .133 bei Testlänge 10 und Dimension vier „Bewerten“ und .503 bei Testlänge 40 und Dimension drei „Integrieren“). Die unter Einsatz des C-MAT I-Algorithmus erreichten Reliabilitäten fallen etwas niedriger aus als die der anderen Bedingungen, in denen multidimensionale adaptive Algorithmen eingesetzt wurden, gleichen sich jedoch mit zunehmender Testlänge an.

7.2.5 Einhaltung von Nebenbedingungen an die Testzusammenstellung – Content Management

Die an die Testzusammenstellung gestellten Nebenbedingungen betrafen die inhaltliche Zusammensetzung des Tests. Demnach sollten Testteilnehmenden aus allen fünf Merkmalsfacetten von ICT-Skills jeweils gleichviele Items vorgelegt werden. Diese Bedingung ist bei den drei Algorithmen mit sequenzieller dimensionsspezifischer Itemauswahl (C-MAT I und III sowie S-UAT) per Definition des Algorithmus gegeben. Für die Algorithmen MAT und CU-MAT wurden hingegen Content Management-Methoden (MPI und MMPI) eingesetzt, um über eine Gewichtung der Iteminformation bei der Itemselektion eine Gleichverteilung der Items entsprechend der fünf Merkmalsdimensionen zu erzielen. Tabelle 7.5 zeigt die prozentuale Anzahl von korrekt administrierten Tests und Items für die beiden Algorithmen, in denen Content

Management-Methoden eingesetzt wurden. Es ist klar zu sehen, dass bei eindimensionaler Itemauswahl die Einhaltung der Nebenbedingungen gelingt (CU-MAT-Algorithmus), während sich bei multidimensionaler Itemauswahl (MAT-Algorithmus) in vielen der administrierten Tests Verletzungen der gesetzten Nebenbedingungen beobachten lassen. Auf Itemebene betrachtet, wird aber auch für den MAT-Algorithmus eine hohe Anzahl an Items korrekt, also entsprechend der Nebenbedingungen administriert.

Tabelle 7.5

Prozentuale Anzahl von Tests und Items, die entsprechend der gesetzten Nebenbedingungen an die Zusammenstellung der adaptiven Tests korrekt administriert wurden.

Testlänge	Testalgorithmen:			
	MAT (mit MMPI)		CU-MAT (mit MPI)	
	Korrekt administrierte Tests	Korrekt administrierte Items	Korrekt administrierte Tests	Korrekt administrierte Items
10	10.33 %	81.68 %	100.00 %	100.00 %
20	2.09 %	85.07 %	100.00 %	100.00 %
30	15.36 %	91.50 %	100.00 %	100.00 %
40	0.06 %	86.02 %	100.00 %	100.00 %

Anmerkungen. MAT = Multidimensional Adaptive Testing; CU-MAT = Combined Unidimensional Multidimensional Adaptive Testing; MMPI = Multidimensional-Maximum-Priority-Index; MPI = Maximum-Priority-Index

7.2.6 Itemvorgabehäufigkeiten – Exposure-Rates

Abbildung 7.3 gibt zunächst einen Eindruck zur Vorgabehäufigkeit der 64 ICT-Items in den jeweiligen Tests. 100 % Administration bedeuten, dass ein Item in jedem der 1000 Tests in 100 Replikationen (insgesamt 100000 Tests) vorgelegt wurde.

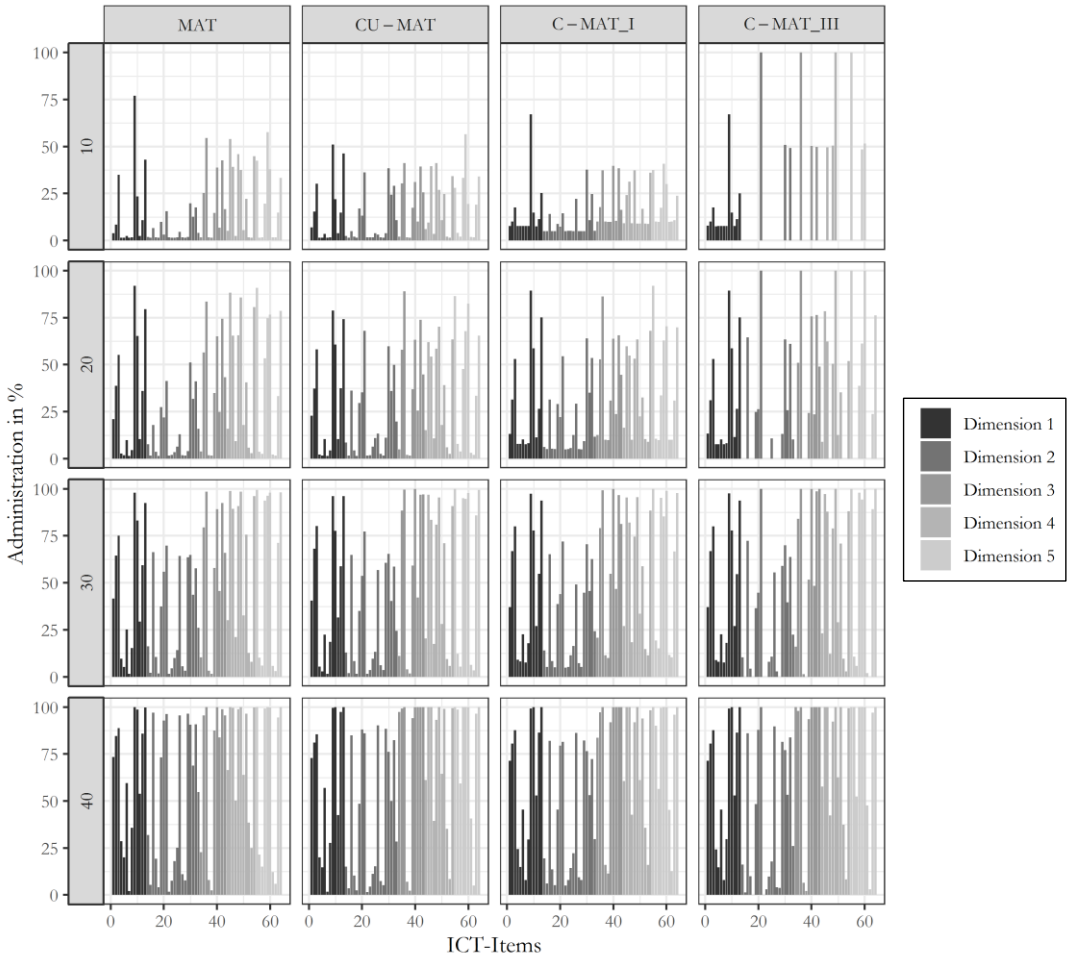

Abbildung 7.3 Vorgabehäufigkeit der 64 ICT-Items nach Testlänge und Testalgorithmus (MAT = Multidimensional Adaptive Testing; C-MAT I & III = Constrained Multidimensional Adaptive Testing I & III)

Die vorgegebenen Items in der C-MAT I- und der S-UAT-Bedingung sind identisch, da sich bei diesen Testalgorithmen nur die abschließende dimensionsspezifische Fähigkeitsschätzung (multidimensional vs. unidimensional) unterscheidet, nicht aber der Prozess der Itemselektion. Daher werden in Abbildung 7.3 und in Tabelle 7.6 nur die Exposure-Rates beziehungsweise TOR für die C-MAT I-Bedingung, nicht aber für die S-UAT-Bedingung dargestellt.

Nur in der C-MAT III-Bedingung gab es für jede der vier betrachteten Testlängen Items, die in keinem der jeweils 100000 Tests administriert wurden. Dabei wurden bei einer Testlänge von 10 Items 39 ICT-Items nicht ausgeliefert, bei einer Testlänge von 20 Items waren es noch 20 nicht-administrierte ICT-Items, bei einer Testlänge von 30 Items wurden 7 ICT-Items keiner einzigen Personen vorgelegt und bei Testlänge von 40 Items wurden immer noch 2 ICT-Items nie administriert.

Diese Beobachtungen zur Itemvorgabehäufigkeit werden durch die TOR, die in Tabelle 7.6 abgetragen sind, bestätigt.

Tabelle 7.6
Über die Replikationen gemittelte Test Overlap Rate (TOR) mit Standardfehler (SE)

Testlänge	Testalgorithmen:			
	MAT	CU-MAT	C-MAT I	C-MAT III
	TOR (*SE*)	*TOR* (*SE*)	*TOR* (*SE*)	*TOR* (*SE*)
10	.374 (.004)	.308 (.003)	.257 (.002)	.663 (.001)
20	.604 (.004)	.558 (004)	.523 (.002)	.663 (.002)
30	.743 (.003)	.761 (.004)	.714 (.002)	.781 (.002)
40	.848 (.002)	.860 (.002)	.827 (.002)	.861 (.001)

Anmerkungen. MAT = Multidimensional Adaptive Testing; CU-MAT = Combined Unidimensional Multidimensional Adaptive Testing; C-MAT I & III = Constrained Multidimensional Adaptive Testing I & III.

Hier zeigen sich vor allem in den C-MAT III-Bedingungen hohe Werte, die sich mit steigender Testlänge denen der anderen Algorithmen annähern. Die für alle Testlängen niedrigsten TOR können mit Einsatz des C-MAT I erreicht werden.

7.2.7 Testzeit

Um eine bessere Einschätzung der zu antizipierenden Testzeit abgeben zu können, wurden anhand der im Rahmen des CavE-ICT-Feldtests beobachteten mittleren Bearbeitungszeiten der einzelnen ICT-Items Testzeiten errechnet. Diese wurden über Personen, Replikationen und Algorithmen gemittelt. Die Bearbeitung eines adaptiven ICT-Tests mit einer Länge von 10 Items würde demnach etwa 17.21 Minuten in Anspruch

nehmen (*SD* = 1.39 Minuten). Bei ICT-Kurztests mit einer Länge von 20 Items würde im Mittel mit einer Testzeit von etwa 35.03 Minuten zu rechnen sein (*SD* = 1.87 Minuten), was sich bei einer Testlänge von 30 ICT-Items auf etwa 52.97 Minuten (*SD* = 1.98 Minuten) und einer Testlänge von 40 Items auf etwa 70.12 Minuten erhöht (*SD* = 1.97 Minuten).

7.3 Diskussion

Im Rahmen der zweiten Studie der vorliegenden Arbeit wurde mithilfe der CavE-ICT-Feldtestdaten exploriert, wie der erarbeitete ICT-Itempool im Rahmen des adaptiven Testens genutzt werden kann. Durch Simulationen wurde untersucht, wie ein adaptiver Test gestaltet sein sollte, der die Facetten des ICT-Skills Konzepts durch Methoden des multidimensionalen adaptiven Testens adressiert und Fähigkeitsschätzungen für Personen für jeden der fünf kognitiven Prozesse ermöglicht. Um Antworten auf die in Kapitel 5.2 formulierten Forschungsfragen geben zu können, wurden neben Messpräzision, -effizienz und Reliabilität auch das Einhalten von Nebenbedingungen an die Testzusammenstellung und die Vorgabehäufigkeit von Items für verschiedene Testalgorithmen analysiert.

Die formulierte Forschungsfrage 2.1 bezieht sich auf die Länge des adaptiven ICT-Skills-Tests. Es zeigt sich für alle simulierten Testalgorithmen die Mehrdimensionalität zur Fähigkeitsschätzung und/oder Itemauswahl einbeziehen, dass die gesetzte Mindestschwelle für Reliabilitätswerte von .55 für den Bericht von Testwerten im Rahmen von LSAs nur für vier der kognitiven Prozesse und erst bei einer Testlänge von insgesamt 40 Items (womit im Idealfall acht Items pro Facette vorgelegt wurden) erreicht werden können. Die Reliabilitäten der einzelnen Dimensionsschätzung unterschieden sich mitunter deutlich. Nur für zwei der fünf Dimensionen („Managen“ und „Integrieren“) konnten bereits bei einer Testlänge von 30 Items Reliabilitäten über

.55 beobachtet werden, hingegen war dies für die Merkmalsfacette „Bewerten“ selbst bei einer Testlänge von 40 Items nicht der Fall. So wie die Reliabilitäten mit steigender Testlänge erwartungsgemäß zunehmen, verringern sich Bias und MSE. Allerdings gibt es auch hier deutliche Unterschiede in den Werten für die verschiedenen kognitiven Prozesse. Die besten Werte sind für die Dimensionen zu beobachten, die mit mehr Items im Pool repräsentiert werden. Besonders gut können die Fähigkeiten des kognitiven Prozesses „Managen“ geschätzt werden. Damit scheint selbst eine Länge des adaptiven Tests von 40 Items noch zu niedrig, um verlässliche Fähigkeitsschätzungen für alle fünf kognitiven Prozesse zu erreichen, was die durch die Ergebnisse aus Studie I empfohlene Itemanzahl mehr als verdoppeln würde.

Verglichen mit den Ergebnissen aus Studie I zeigt sich in den Resultaten der vorliegenden Studie über die verschiedenen Versuchsbedingungen hinweg ein etwas höherer Bias, vor allem mit Blick auf einzelne Dimensionen. Hingegen fallen die MSEs etwas niedriger aus als auf Basis der Ergebnisse von Studie I erwartet. Die Einschätzungen hinsichtlich der relativen Messeffizienz konnten dahingehend bestätigt werden, dass Vorteile aller multidimensionalen Tests im Vergleich zum Testalgorithmus mit unidimensionaler Itemauswahl und wiederholter unidimensionaler Fähigkeitsschätzung für die fünf Merkmalsdimensionen bzw. -facettenebenen gefunden wurden. In Bezug auf die Reliabilitäten der Fähigkeitsschätzungen muss allerdings konstatiert werden, dass diese deutlich geringer ausfallen als auf Basis der vorhergehenden Simulationsstudie erwartet wurde. Dies ist wahrscheinlich dem geschuldet, dass die durch Studie I empfohlene Anzahl an Items im ICT-Pool nicht ganz erreicht wurde und sich diese Items zudem nicht auf die fünf kognitiven Prozesse und über das Fähigkeitsspektrum der Zielpopulation hinweg gleich verteilen. Des Weiteren fallen auch die im Zuge des CavE-ICT-Feldtests beobachteten Korrelationen zwischen den Merkmalsfacetten, die durch die verschiedenen Testalgorithmen genutzt wurden, sehr unterschiedlich aus. So liegt die im Rahmen des

CavE-ICT-Feldtests beobachtete Korrelation zwischen den kognitiven Prozessen „Zugreifen“ und „Managen“ mit .72 leicht über den in Studie I antizipierten .70. Hingegen konnte zwischen den kognitiven Prozessen „Zugreifen“ und „Bewerten“ nur ein Zusammenhang von .48 beobachtet werden, was wiederum leicht unter den in Studie I einbezogenen von .50 liegt. Es ist zu schließen, dass diese starken Abweichungen zwischen im CavE-ICT-Feldtest beobachteten und im Rahmen von Studie I antizipierten Werten zu einer Verschlechterung der Simulationsergebnisse führen.

Die Forschungsfragen 2.2 und 2.3 beziehen sich auf einen Vergleich des MAT-Algorithmus mit eingeschränkten multidimensionalen Testalgorithmen sowie mit einem Algorithmus, der unidimensionale Itemauswahl mit multidimensionaler Fähigkeitsschätzung kombiniert. MAT schneidet generell etwas besser ab als die eingeschränkten Algorithmen, wobei MAT und CU-MAT ähnlich gut funktionieren und bei einigen Kriterien sogar leichte Vorteile des CU-MAT zu finden sind. Bias und MSE aller multidimensionalen Algorithmen fallen ähnlich hoch aus, wobei C-MAT I am schlechtesten abschneidet. Mit zunehmender Testlänge gleichen sich die Werte aber noch stärker an. Auch im Hinblick auf die RE schneidet der C-MAT I am schlechtesten ab, liefert allerdings nicht deutlich schlechtere Resultate als die anderen multidimensionalen Algorithmen.

Die Ergebnisse sprechen dafür, dass es durchaus Alternativen zum Einsatz von MAT gibt, wenn eine dimensionsspezifische Itemvorgabe erwünscht ist. Gegeben des in dieser Studie genutzten eher nicht optimalen und kleinen Itempools (verglichen mit den im Rahmen von Studie I simulierten Pools) sind die Einbußen im Hinblick auf Messgenauigkeit, -effizienz und Reliabilität bei Nutzung eines Algorithmus mit sequentieller dimensionsspezifischer Itemvorgabe und multidimensionaler abschließender Fähigkeitsschätzung nur gering. Daraus lässt sich schließen, dass die restriktiveren C-MAT-Algorithmen bei ausgewogeneren und etwas größeren Itempools noch stärker als

bei Verwendung des vorliegenden ICT-Itempools als Alternativen zum MAT in Betracht gezogen werden können.

Mit Blick auf das Content-Management muss festgehalten werden, dass der MMPI bei Einsatz des MAT-Algorithmus nicht optimal funktioniert. Allerdings stehen nicht allzu viele Items für eine große Auswahl im Pool zur Verfügung. Bei einer Testlänge von 40 Items müssten entsprechend der gesetzten Nebenbedingungen an die Testzusammenstellung 8 Items zu jedem der fünf kognitiven Prozesse ausgewählt werden, es stehen jeweils aber nur 13, 20, 10, 11 und 10 Items zur Verfügung (vgl. Tabelle 4.2 in Kapitel 4). Für alle anderen Algorithmen konnten die Nebenbedingungen bezüglich der Vorgabehäufigkeit von Items zur Messung der fünf kognitiven Prozesse eingehalten werden.

In Bezug auf die Item-Exposure-Rates bei Nutzung der verschiedenen Algorithmen ist festzustellen, dass der C-MAT III-Algorithmus den zur Verfügung stehenden Itempool nicht optimal ausnutzt, vor allem wenn die Testlänge gering ist. Doch selbst bei 40 vorgelegten ICT-Items gibt es zwei Items die bei Einsatz dieses Algorithmus in keinem der 1000*100 Tests vorgelegt wurden. Hingegen nutzt der C-MAT I-Algorithmus den Itempool am besten aus. Wahrscheinlich gelingt dies durch die mehrfache zufällige Auswahl von Items im Testverlauf, was auf der anderen Seite allerdings auch zu Lasten der Messgenauigkeit und Reliabilität geht. Unterstützt wird dieses Ergebnis durch die Betrachtung der TOR, die wiederum für die C-MAT I-Bedingungen am niedrigsten ausfällt und für eine vergleichsweise geringere Überschneidung von Testitems zwischen Personen steht. Insgesamt betrachtet steigt die TOR jedoch für alle Testalgorithmen mit steigender Testlänge, was mit dem verhältnismäßig kleinen Itempool zusammenhängt, da bei einer Testlänge von 40 Items fast zwei Drittel des Itempools vorgelegt werden müssen.

Sollen Testitems dimensionsspezifisch sequentiell administriert werden, bietet sich, unter Berücksichtigung aller zuvor betrachteten Kriterien, zum Einsatz am ehesten der

C-MAT I-Algorithmus an. Zwar fallen die in dieser Bedingung beobachteten Werte für MSE, RE und Reliabilität geringfügig schlechter als für den C-MAT III aus, allerdings wird unter Einsatz dieses Algorithmus der Itempool sehr viel besser ausgenutzt. Zusammenfassend muss allerdings festgehalten werden, dass selbst eine Testlänge von 40 Items nicht hinreichend ist, um ausreichend reliable Fähigkeitsschätzungen für alle fünf kognitiven Prozesse erreichen zu können. Zudem steigt auch die zu veranschlagende Testzeit, die bei einer Testlänge von 40 Items durchschnittlich schon bei über 70 Minuten läge.

Da die Testlänge und damit verbunden die zu antizipierende Testzeit so hoch und letztlich recht viele der Items aus dem Pool vorgegeben werden müssten, steht in Frage, ob sich die diagnostischen Vorteile der nach kognitiven Prozessen differenzierten Rückmeldung letztlich rentiert. Zum Vergleich konnte unter Einsatz eines eindimensionalen adaptiven Tests, der eine Gleichverteilung der Items entsprechend der kognitiven Prozesse über eine Content Management Methode sicher stellte, dann allerdings „nur“ eine ICT-Gesamtfähigkeit ermittelt, bereits nach Vorlage von 17 Items eine Reliabilität von .55 erreicht werden (vgl. Kapitel 4.3.3). Wie unter Nutzung des in dieser Studie eingesetzten CU-MAT-Algorithmus wurden die Items dabei auch nicht geordnet nach kognitivem Prozess sondern gemischt vorgegeben.

Eine naheliegende weitere Überlegung ist es daher, eine Auswahl von ICT-Items vorzunehmen, welche die Rahmenkonzeption von ICT-Skills (vgl. Kapitel 4.2) möglichst gut abbildet und als linearer Kurztest mit festgesetzter Itemreihenfolge vorgelegt werden kann. In der dritten Studie dieser Arbeit werden dementsprechend verschieden lange ICT-Kurztests zusammengestellt und durch Simulationsstudien eruiert ob diese im Kontext von LSAs oder zur Individualdiagnostik eingesetzt werden könnten.

8 Studie III – Zusammenstellung von Kurztests zur eindimensionalen Erfassung von ICT-Skills unter Nutzung der im Zuge des CavE-ICT-Feldtests geschätzten Itemparameter

Im folgenden Kapitel werden Methode, Ergebnisse und Diskussion der dritten empirischen Studie dieser Arbeit dargestellt. Im Rahmen dieser Studie werden, wie in Kapitel 5.3 ausgeführt, lineare Kurztests zur Erfassung von ICT-Skills zusammengestellt und evaluiert. Dabei werden zwei unterschiedliche Ansätze der Testzusammenstellung verfolgt und verglichen. Im Gegensatz zu den zuvor beschriebenen Studien liegt der Fokus nachfolgend auf der Exploration der Möglichkeiten zur Erfassung des Gesamtkonstrukts ICT-Skills über einen eindimensionalen und nicht-adaptiven Test. Zudem soll im Zuge dieser Studie der Einsatz eines solchen Tests nicht nur im Kontext von LSAs diskutiert werden, sondern auch die Nutzbarkeit im Rahmen von Untersuchungen kleinerer Gruppen, zu Screening-Zwecken beziehungsweise im Bereich der Individualdiagnostik erwogen werden.

8.1 Methode

Die in Kapitel 5.3 dargestellten Forschungsfragen wurden bearbeitet, indem zunächst aus dem ICT-Itempool verschiedene Kurztests zusammengestellt und anschließend deren Performanz mithilfe von Simulationsstudien analysiert wurde. Die Zusammenstellung von ICT-Kurztests erfolgte auf zwei unterschiedliche Arten und wird zunächst dargestellt. Das zur vergleichenden Evaluation der so erstellten Tests genutzte Studiendesign, die Datengenerierung und die herangezogenen Evaluationskriterien werden anschließend erläutert. Ziel dieser Studie war es herauszufinden wie lang ein linearer, möglichst optimal

zusammengestellter ICT-Kurztest, der verschiedene Nebenbedingungen erfüllt, mindestens sein muss, um ICT-Skills in Bezug auf die aufgewendete Testzeit ökonomisch sowie messeffizient, präzise und zuverlässig zu erfassen. Zudem sollte der ICT-Kurztest nicht nur auf Populationsebene im Kontext von LSAs sondern auch im Rahmen von Gruppen- und Individualdiagnostik einsetzbar sein.

8.1.1 Zusammenstellung der Kurztests

Vorgehen bei der Testzusammenstellung. Um die oben genannten Ziele zu erreichen wurden zunächst unterschiedliche ICT-Kurztests zusammengestellt. Die Zielpopulation der Tests setzt sich aus 15-jährigen Schülerinnen und Schülern zusammen. Die Stichprobe des CavE-ICT-Feldtests wurde aus dieser Population gezogen, so dass die aus diesen Daten geschätzte ICT-Fähigkeitsverteilung (siehe Abbildung 4.3 in Kapitel 4) zur Orientierung bei der Itemselektion für die Kurztests genutzt wurde. Das durch die Itemauswahl abgedeckte Schwierigkeitsspektrum sollte in diesem Sinne weitestgehend der Fähigkeitsverteilung der Zielpopulation entsprechen. Zusätzlich zu dieser Zielformulierung wurden auch Nebenbedingungen aufgestellt, die in der Testzusammenstellung berücksichtigt werden sollten. Insgesamt sollen vier Kurztests unterschiedlicher Länge und unterschiedlich langer zu approximierender Testzeit zusammengestellt werden (10, 15, 20 und 25 Items mit maximal 20, 30, 40 und 50 Minuten geschätzter mittlerer Testzeit), wobei zuvor selektierte Items der kürzeren Tests in den längeren Tests enthalten sein sollten. Die Komplexität des zugrundeliegenden theoretischen Rahmenkonzepts zur Beschreibung von ICT-Skills (vgl. Abbildung 4.2 in Kapitel 4) sollte durch die ausgewählten Items ebenso abgedeckt werden wie die Vielfalt gängiger Computeranwendungen. Schließlich sollten die selektierten Items hinreichend hoch diskriminieren. Gegeben der Eigenschaften des ICT-Itempools wird in diesem Zusammenhang eine Korrelation des Items mit dem Gesamttestwert von mindestens .30 als ausreichend angesehen um es für die Kurztests zu selektieren (Bortz und Döring, 2006).

Eine Übersicht der Zielsetzungen und Nebenbedingungen der Testzusammenstellung wird nachfolgend in Tabelle 8.1 gegeben.

Tabelle 8.1

Zielsetzung und Nebenbedingungen für die Zusammenstellung von ICT-Kurztests aus den zur Verfügung stehenden 64 ICT-Items

Zielsetzung (Objective)	Das durch die Itemauswahl abgedeckte Schwierigkeitsspektrum sollte weitestgehend auch der Fähigkeitsverteilung der Zielpopulation entsprechen.	
Neben-bedingungen (Constraints)	1	**Testlänge**: Es wurden jeweils Tests mit einer Länge von 10, 15, 20 und 25 Items zusammengestellt.
	2	**Aufbauende Selektion:** Die zuvor selektierten Items der kürzeren Tests sollten jeweils die Basis der längeren Tests darstellen.
	3	**Testzeit**: Es sollten durchschnittlich 10 Minuten für die Bearbeitung von 5 Items nicht überschritten werden. Daher ergibt sich eine maximale Testzeit von 20, 30, 40 bzw. 50 Minuten für die jeweiligen Tests.
	4	**Rahmenkonzeption:** Die theoretische Rahmenkonzeption zur Beschreibung von ICT-Skills sollte möglichst umfassend durch die ausgewählten Items repräsentiert werden. Genau $1/5$ der Items in den jeweiligen Tests sollten die kognitiven Prozesse abdecken. Mindestens $1/5$ der Items in den jeweiligen Tests sollten die Ebenen der Facette Situation und mindestens $1/3$ der Items in den jeweiligen Tests die Ebenen der Facette soziale Interaktion repräsentieren.
	5	**Applikationen**: Es sollte durch die Items eine möglichst große Spannbreite gängiger computerbasierter Anwendungen abgedeckt werden, indem aus jeder im ICT-Itempool wiederholt dargestellten Applikation mindestens ein Item im Kurztest sein sollte.
	6	**Itemdiskrimination**: Die ausgewählten Items sollten hinreichend hoch diskriminieren ($r_{pbis} \geq .30$).

Im Rahmen von Studie III wurden unter Berücksichtigung der vorangegangenen Überlegungen zwei Ansätze zur Testzusammenstellung verfolgt: zum einen wurden Tests manuell, zum anderen automatisiert computerbasiert (unter Verwendung linearer Programmierung; van der Linden, 1998b, 2005) erstellt. Für die *manuelle Testzusammenstellung* wurde versucht Items für die Kurztests so auszuwählen, dass die empirische Verteilung der Personenfähigkeiten des CavE-ICT-Feldtests möglichst umfassend durch Items unterfüttert wird. Ziel war es in jedem Bereich der Fähigkeitsverteilung auch Items zur Differenzierung von Personen vorliegen zu haben.

Parallel wurde darauf geachtet in den jeweiligen Kurztests die in Tabelle 8.1 dargestellten Nebenbedingungen einzuhalten. Hohe Priorität wurde der Abdeckung der theoretischen Rahmenkonzeption gegeben. Wie in den vorherigen Studien wurde eine Gleichverteilung von ICT-Items auf die fünf kognitiven Prozesse in den jeweiligen Tests fokussiert. Zusätzlich sollten aber auch die Ebenen der Facetten Soziale Interaktion (individuell und kollektiv) und Situation (persönlich, beruflich und bildungsbezogen) durch hinreichend viele Items in den ICT-Kurztests repräsentiert werden. Zudem wurden Items bevorzugt ausgewählt, die eine möglichst geringe mittlere Bearbeitungszeit im CavE-ICT-Feldtest aufwiesen.

Van der Linden (2005) beschreibt die Testzusammenstellung als Problem der eingeschränkten kombinatorischen Optimierung, welches mit Hilfe linearer Programmierung computerbasiert gelöst werden kann. Bei einer in dieser Weise *automatisierten Testzusammenstellung* gilt es aus einem vorgegebenen Itempool die Items beziehungsweise den Test zu selektieren, der zum einen die formulierten Testziele erfüllt, dabei aber auch zuvor definierte Nebenbedingungen einhält. Die von van der Linden (2005) angestellten Überlegungen zur Formulierung eines mathematischen Optimierungsproblems sollen im Folgenden kurz und nur für im Rahmen dieser Studie als relevant erachtete Aspekte dargestellt werden (für eine ausführliche Beschreibung siehe van der Linden, 2005). Anschließend werden konkret die für diese Studie formulierten Nebenbedingungen und die Zielfunktion mathematisch dargestellt.

Ein Optimierungsproblem, bei dem aus einer Menge von Items $i = 1,...,I$ ein Kurztest abgeleitet werden soll, hat I Entscheidungsvariablen x_i, die jeweils definiert sind als

$$x_i = \begin{cases} 1 & \text{wenn Item } i \text{ für den Test selktiert wurde} \\ 0 & \text{wenn Item } i \text{ nicht für den Test selektiert wurde.} \end{cases} \tag{8.1}$$

Über diese Entscheidungsvariablen ist es möglich, Nebenbedingungen für die Testzusammenstellung zu definieren. Dabei wird unter Berücksichtigung dieser Einschränkungen die Auswahl aus der Gesamtheit aller möglichen Tests auf die Auswahl aus der Teilmenge der durchführbaren Tests reduziert. Nebenbedingungen können quantitativer, kategorialer und logischer Art sein. Quantitative Nebenbedingungen werden über Merkmale definiert, die numerische Werte annehmen und haben reelle Grenzen für gewichtete Variablensummen. Kategoriale Nebenbedingungen haben hingegen ganzzahlige Grenzen für ungewichtete Summen von Variablen und werden über Merkmale definiert, die den Itempool in Teilmengen mit einem gemeinsamen Merkmal gliedern. Da die Entscheidungsvariablen binär sind, kann die Auswahl einer bestimmten Anzahl von Items aus einer gegebenen Teilmenge realisiert werden, indem Grenzen für einfache Variablensummen gesetzt werden. Quantitative und kategoriale Nebenbedingungen können auf Test-, Itemset- oder Itemebene definiert werden. Logische Nebenbedingungen implizieren eine Wenn-Dann-Auswahl oder eine bedingte Auswahl von Elementen. Logische Nebenbedingungen können über Ausschluss (das heißt wenn ein Item selektiert wurde, können bestimmte andere Items nicht mehr im Test sein) oder Einbezug (das heißt, wenn ein Item selektiert wurde, müssen auch bestimmte andere Items im Test sein) definiert werden.

Unter Einhaltung solcher Nebenbedingungen wird die Testinformationsfunktion (TIF) des zusammengestellten Tests im Hinblick auf eine zuvor definierte Zielsetzung optimiert werden. Dabei wird darauf zurückgegriffen, dass die einzelnen Iteminformationskurven additiv sind. Die Optimierung der Testzusammenstellung erfolgt mit dem Ziel, die resultierende TIF an eine vorgegebene Form anzupassen. Die Funktion $R(\theta)$ an deren Form die TIF angepasst werden soll wird über Zielwerte $R_k = R(\theta_k), k = 1,...,K$ entlang der θ-Skala definiert. Dabei sind zwischen drei und fünf Zielwerte ausreichend, um eine gute Passung von TIF und $R(\theta)$ zu erreichen (van der

Linden, 2005). Da für mehrere Zielwerte entlang der θ-Skala gleichzeitig optimiert werden muss, spricht van der Linden (2005) von Multiobjective-Test-Assembly. Das relative Ziel wird über Werte $R_k > 0$ definiert, die die nötige Information am Punkt θ_k relativ zu den anderen Werten $k = 1,...,K$ repräsentieren. Die TIF ist dann an allen Punkten θ_k, $k = 1,...,K$ zu maximieren, wobei gleichzeitig die relative Form der Zielfunktion beibehalten wird. Da die Zielwerte keine feste Einheit haben, kann ein Wert auf eins gesetzt und die übrigen Werte entsprechend angepasst werden. Wird $R_1 = 1$ gesetzt, folgt daraus, dass für $K \geq 2$ am Punkt θ_k die TIF R_k mal größer sein muss als an θ_1:

$$\sum_{i=1}^{I} I_i(\theta_k)x_i = R_k \sum_{i=1}^{I} I_i(\theta_1)x_i \text{ , für } k \geq 2 . \tag{8.2}$$

Um die TIF entlang der K Werte θ_k gleichzeitig zu maximieren, muss so nur noch die TIF an einem dieser Werte maximiert werden. Die Formalisierung könnte dann folgendermaßen aussehen:

$$\text{maximiere } \sum_{i=1}^{I} I_i(\theta_1)x_i . \tag{8.3}$$

Da es eher unwahrscheinlich ist, dass Gleichung 8.2 genau erfüllt werden kann, empfiehlt van der Linden (2005) für diesen Fall Ungleichheit zuzulassen und so die Realisierbarkeit der Bedingung zu gewährleisten:

$$\sum_{i=1}^{I} I_i(\theta_k)x_i \geq R_k \sum_{i=1}^{I} I_i(\theta_1)x_i \text{ , für } k \geq 2 . \tag{8.4}$$

Um zu einem Kriterium zu gelangen, dass alle R_k gleichförmig berücksichtigt, schlägt van der Linden (2005) vor, den Ausdruck $\sum_{i=1}^{I} I_i(\theta_1)x_i$ durch eine neue Variable y zu ersetzen. Dadurch ergibt sich schließlich als Zielsetzung für die Testzusammenstellung:

$$\text{maximiere } y \tag{8.5}$$

unter der Bedingung

$$\sum_{i=1}^{I} I_i(\theta_k) x_i \geq R_k y \text{, für alle } k. \tag{8.6}$$

Mit so formulierten Nebenbedingungen und Zielsetzungen für die Testzusammenstellung kann ein Modell zur Lösung des Problems beziehungsweise für die Auswahl eines geeigneten Tests erstellt werden. Nebenbedingungen und Zielsetzung sind als lineare Gleichungen beziehungsweise Ungleichungen ausgedrückt, so dass dieses Modell ein Beispiel für ein ganzzahliges lineares 0-1 Programmierproblem ist. Eine Lösung für dieses Modell ist ein Wertevektor für die 0-1-Variablen $(x_{i=1},...,x_I)$, der alle Bedingungen erfüllt und die Zielfunktion optimiert.

Die konkret im Rahmen dieser Studie formulierten Nebenbedingungen beziehen sich auf den CavE-ICT Itempool mit $I = 64$ Items. Die in Tabelle 8.1 bereits benannten Nebenbedingungen wurden als (Un-)Gleichungen formuliert, wobei sie so festgehalten wurden, dass sie für jeden der vier zusammenzustellenden Kurztests der Länge m_l mit $l = 1,...,4$ gelesen werden können. Daher werden die Entscheidungsvariablen x_{il} doppelt indiziert, was neben der gemeinsamen Spezifikation von Items und Tests, auch der Formalisierung der Inklusion der Items aus den Tests $l-1$ und damit der sukzessiven Testzusammenstellung dient.

Die aus den CavE-ICT-Feldtestdaten geschätzte mittlere Bearbeitungszeit eines Items i wird mit t_i bezeichnet, die Diskrimination mit $r_{\text{pbis}\,i}$. Die Menge der Items im Pool, die eine der fünf kognitiven Prozesse $h \in \{1,...,5\}$ zur Bearbeitung erfordert, wird mit V_h^c bezeichnet. Mit V_h^s wird auf die Menge von Items referiert, die eine der drei Situationen $h \in \{1,...,3\}$ abbildet, und mit V_h^z wird die Menge der Items bezeichnet, die eine der beiden sozialen Interaktionen $h \in \{1,2\}$ erfordert. Mit V_h^a wird die Menge der

Items bezeichnet, die eine der fünf Kategorien von Computerapplikationen $h \in \{1,...,5\}$ darstellen. Item $i \in V_h^a$ ist somit ein beliebiges Item im Pool, welches Applikation h darstellt.

Zur Definition der Zielfunktion dient wiederum die empirische Verteilung der Personenfähigkeiten des CavE-ICT-Feldtests. Die Testinformationskurven der einzelnen Kurztests sollten der relativen Zielfunktion $R(\theta)$ entsprechen, die im Rahmen dieser Studie durch Zielwerte an den Punkten $(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5) = (-1.5,-1,0,1,1.5)$ auf Basis der empirischen Verteilung der Personenfähigkeiten des CavE-ICT-Feldtests (vgl. Abbildung 4.3 in Kapitel 4 oder im Folgenden Abbildung 8.1) mit $(R_1,R_2,R_3,R_4,R_5) = (1.44,7.18,24.54,19.58,4.96)$ definiert ist.

Als Zielsetzung für die Zusammenstellung beziehungsweise Auswahl von Kurztests ergibt sich daher, in leichter Abwandlung zu Formel 8.6:

$$\text{maximiere } y \tag{8.7}$$

unter der Bedingung, dass

$$\sum_{i=1}^{I} I_i(\theta_k) x_{il} \geq R_k y \text{, für alle } k. \tag{8.8}$$

Dabei sind folgende Nebenbedingungen einzuhalten:

$$\sum_{i=1}^{I} x_{il} = m_l \qquad \text{(Testlänge)} \tag{8.9}$$

$$x_{il} = x_{i(l-1)} \qquad l \geq 2 \qquad i : x_{i(l-1)} = 1 \qquad \text{(Aufbauende Selektion)} \tag{8.10}$$

$$\sum_{i=1}^{I} t_i x_{il} \leq 2m_l \qquad \text{(Testzeit)} \tag{8.11}$$

$$\sum_{i \in V_h^i} x_{il} \geq \frac{m_l}{5} \qquad h = 1,...,5 \qquad \text{(Kognitive Prozesse)} \tag{8.12}$$

$$\sum_{i \in V_h^s} x_{il} \geq \frac{m_l}{5} \qquad h = 1,...,3 \qquad \text{(Situation)} \qquad (8.13)$$

$$\sum_{i \in V_h^z} x_{il} \geq \frac{m_l}{3} \qquad h = 1,2 \qquad \text{(Soziale Interaktion)} \qquad (8.14)$$

$$\sum_{i \in V_h^a} x_{il} \geq 1 \qquad h = 1,...,5 \qquad \text{(Computerapplikationen)} \qquad (8.15)$$

$$r_{\text{pbis } i} \geq .30 x_{il} \qquad \text{(Itemdiskrimination)} \qquad (8.16)$$

Die vier ICT-Kurztests wurden gegeben dieser Zielsetzungen und Nebenbedingungen sukzessive für $l = 1,...,L$ erstellt, wobei zuvor selektierte Items des Tests $l-1$ für den Test l gesetzt sind. Verwendet wurde das Programm GLPK (GNU Linear Programming Kit; Andrew Makhorin, 2012). Dieses greift auf einen Branch-and-Bound-Algorithmus zurück, der auf einer impliziten Aufzählung aller möglichen Lösungen (also aller möglichen Testzusammenstellungen) für ein Problem basiert. In einem iterativen Prozess wird der gesamte Lösungsraum nach einem Optimum durchsucht, wobei Teile des Raums, für die ein Optimum unmöglich erscheint direkt verworfen werden, was die Lösung des Optimierungsproblems praktisch möglich und den Ansatz effizient macht (van der Linden, 2005). Der Algorithmus stoppt, sobald keine Verbesserung einer bereits gefundenen Lösung mehr möglich ist. Damit resultiert ein ICT-Kurztest der unter Einhaltung der oben beschriebenen Zielsetzungen und Nebenbedingungen optimal zusammengestellt ist.

Resultierende Kurztests. Abbildung 8.1 stellt die Verteilungen von Itemschwierigkeiten für die vier manuell und die vier automatisiert zusammengestellten ICT-Kurztests mit der Verteilung der geschätzten Personenfähigkeiten aus dem CavE-ICT-Feldtest dar. Im manuell und automatisiert erstellten Test mit einer Testlänge von 10 Items stimmen nur drei Items überein, für die beiden Tests mit 15 Items sind es fünf.

Zehn Items sind in den 20 Items umfassenden Tests gleich und schließlich 18 Items wurden sowohl manuell als auch automatisiert für die längsten zusammengestellten Tests mit insgesamt 25 Items ausgewählt. Eine detaillierte Übersicht der für die Kurztests selektierten Items kann Anhang D Tabelle D.1 und D.2 entnommen werden.

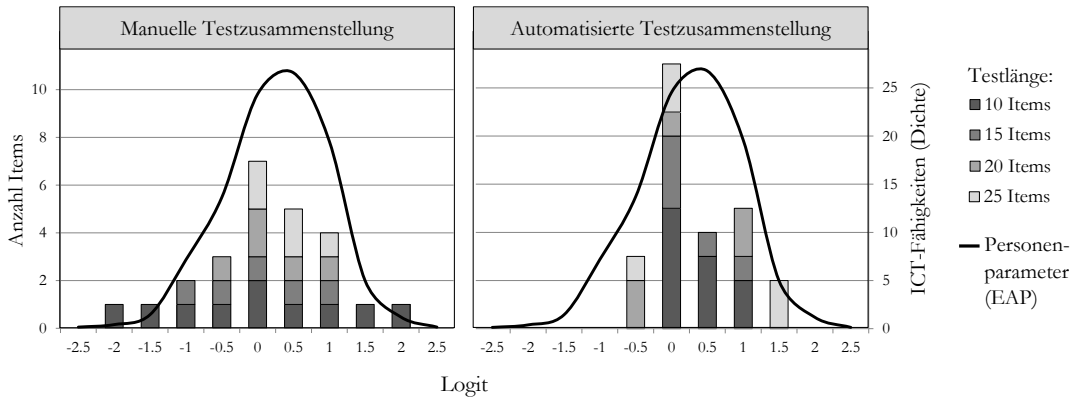

Abbildung 8.1 Verteilung der Schwierigkeiten der für die verschiedenen Kurztests selektierten ICT-Items (Balken) sowie die im CavE-ICT-Feldtest geschätzte Verteilung der ICT-Personenfähigkeiten (Linie).

Es fällt auf, dass bei der manuell erfolgten Testzusammenstellung ein breiterer Bereich der Fähigkeitsverteilung durch die unterschiedlich langen Kurztests abgedeckt wird. Die Itemschwierigkeiten liegen für alle vier manuell zusammengestellten Kurztests zwischen -1.885 und 1.886, wobei die Standardabweichungen der Itemschwierigkeiten mit zunehmender Testlänge abnehmen. Für die automatisiert zusammengestellten Tests zeigt sich ein anderes Bild: Die Itemschwierigkeiten liegen hier für die Kurztest mit 10 Items zwischen -0.196 und 0.819, mit 15 Items zwischen -0.196 und 1.012, für 20 Items Testlänge zwischen -0.697 und 1.012 sowie schließlich für den 25 Items umfassenden Test zwischen -0.697 und 1.571. Die Standardabweichung der Itemschwierigkeiten nimmt mit der Testlänge zu, fällt aber insgesamt niedriger aus als für die manuell zusammengestellten Tests (Tabelle 8.2). Zudem ist die Verteilung der Itemschwierigkeiten der manuell

zusammengestellten Kurztests im Vergleich zu den automatisiert zusammengestellten Tests nur vergleichsweise leicht rechtsschief.

Betrachtet man zusätzlich zur Verteilung der Itemschwierigkeiten auch die jeweiligen Testinformationskurven (als Summe der einzelnen Iteminformationskurven, vgl. Kapitel 2.1.1) in Abbildung 8.2 zeigt sich deutlicher, dass durch die automatisiert zusammengestellten Tests die angestrebte Form der TIF sehr gut abgebildet wird. Die Items der manuell erstellen Tests decken eher einen breiteren als den über die Zielsetzung definierten erforderlichen Bereich ab.

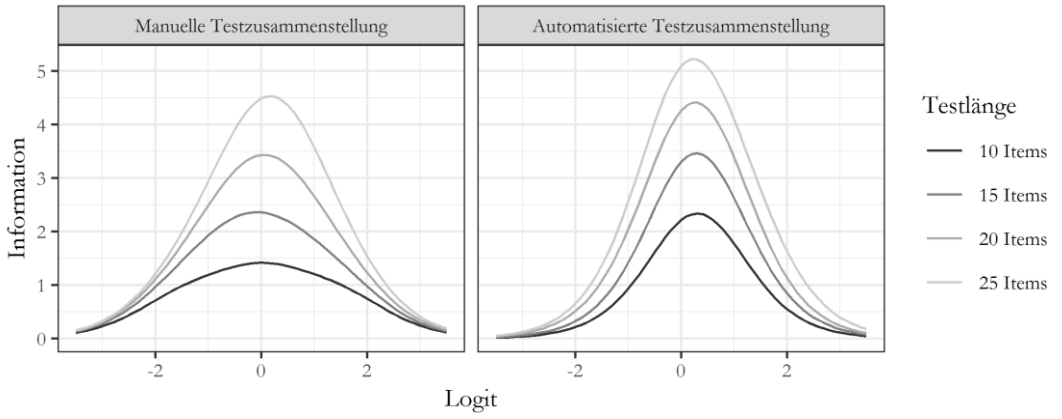

Abbildung 8.2 Testinformationskurven der verschiedenen ICT-Kurztests.

Insgesamt lässt sich eine recht gute Passung zwischen Itemschwierigkeiten und der Verteilung der Personenfähigkeiten für alle zusammengestellten ICT-Kurztests konstatieren. Vor allem im mittleren Fähigkeitsbereich sollten in jedem der Kurztests hinreichend viele Items zur Verfügung stehen, um Differenzierbarkeit zwischen Personen zu ermöglichen. Im Hinblick auf die psychometrische Güte der für die Kurztests ausgewählten Items werden zur Beurteilung des Itemfit die Werte des WMNSQ (Weighted-Mean-Square; gewichtetes Abweichungsquadrat; vgl. Wright & Linacre, 1994) betrachtet, welche einen Erwartungswert von 1 haben und zwischen 0.8 und 1.2 liegen sollten (Bond & Fox, 2007). Wie Tabelle 8.2 entnommen werden kann, liegen die Werte

des WMNSQ aller ausgewählten Items in diesen Grenzen, was für eine gute Passung der Items zum Rasch-Modell spricht (Bond & Fox, 2007). Die korrespondierenden *t*-Werte fallen nicht größer als 1.96 aus, was dafür spricht, dass von 1 abweichende Fitwerte der ausgewählten Items statistisch zufällig und nicht systematisch vorliegen. Eine systematische Abweichung kann als Hinweis auf eine zu niedrige Trennschärfe gewertet, wohingegen negative t-Werte auf eine zu hohe Trennschärfe hinweisen und in der Regel als nicht problematisch angesehen werden (Bond & Fox, 2007).

Tabelle 8.2

Zusammenschau und Vergleich der verschieden ICT-Kurztests hinsichtlich Itemschwierigkeit, -fit und -diskrimination

Kurztest-zusammenstellung	Test-länge	Item-schwierigkeit	Itemfit					Itemdiskrimination		
			WMNSQ			*t*-Wert				
		M (SD)	Min	Max	*M*	Min	Max	Min	Max	*M*
Manuell	10	0.024 (1.238)	0.88	0.99	0.93	-2.1	-0.1	.33	.52	.45
	15	0.001 (1.071)	0.88	1.04	0.95	-2.1	0.9	.31	.52	.43
	20	0.048 (0.952)	0.88	1.04	0.96	-2.5	0.9	.31	.55	.43
	25	0.111 (0.874)	0.84	1.09	0.96	-4.3	1.9	.31	.56	.43
Automatisiert	10	0.307 (0.331)	0.89	1.09	1.01	-2.5	1.9	.31	.55	.42
	15	0.307 (0.356)	0.89	1.09	0.99	-2.5	1.9	.31	.55	.43
	20	0.268 (0.455)	0.89	1.09	0.99	-2.5	1.9	.31	.55	.42
	25	0.312 (0.578)	0.89	1.09	0.99	-2.5	1.9	.31	.55	.41

Anmerkung. WMNSQ = Weighted Mean Square (gewichtetes Abweichungsquadrat)

Es zeigen sich des Weiteren hohe Korrelationen der Items mit dem Gesamttestwert, welche im Sinne einer Itemdiskrimination interpretiert werden können. Die Werte liegen für die einzelnen Tests im Mittel zwischen .43 und .45 bei manueller, und zwischen .41 und .43 bei automatisierter Testzusammenstellung.

Im Folgenden wird beschrieben, wie die weiteren Nebenbedingungen der Testzusammenstellung realisiert wurden. Die verschiedenen Facettenebenen sollten durch die ICT-Kurztests möglichst gut abgedeckt werden, um eine inhaltsvalide Interpretation von Testwerten zu ermöglichen. Es wird jedoch nicht angestrebt, durch die Kurztests spezifische Rückmeldungen zu den verschiedenen Facettenebenen zu geben; vielmehr soll die ICT-Fähigkeit trotz ihrer inhaltlichen Komplexität mit einem eindimensionalen Test

erfasst werden. Durch die für die Kurztests jeweils ausgewählten ICT-Items wurden die fünf kognitiven Prozesse (Zugreifen, Managen, Integrieren, Bewerten und Erzeugen) gleichverteilt abgedeckt. Zudem wurden die Tests so zusammengestellt, dass jeweils Items aus bildungsbezogenen, beruflichen und privaten Kontexten vorlagen, womit auch alle Facettenebenen der Situation abgebildet wurden. Die Facette der Sozialen Interaktion wurde ebenso mit Items, die jeweils die individuelle oder kollektive Ebene erfassen, in den Tests repräsentiert. Damit wurde die Komplexität des theoretisch formulierten und bereits in Kapitel 4.2 vorgestellten ICT-Konstrukts in der Itemauswahl für die ICT-Kurztests deutlich abgebildet. Tabelle 8.3 zeigt die Abdeckung der in Kapitel 4 vorgestellten theoretischen Rahmenkonzeption zur Beschreibung von ICT-Skills an den konkreten Itemzahlen. Beim Vergleich der Itemzahlen nach Facetten und Facettenebenen der ICT-Skills Rahmenkonzeption der manuell und der automatisiert zusammengestellten Tests werden trotz Auswahl unterschiedlicher Items nur leichte Unterschiede in der Abdeckung der Rahmenkonzeption deutlich.

Tabelle 8.3

Abdeckung der durch die ICT-Skills Rahmenkonzeption definierten Facetten Situation und soziale Interaktion

Kurztest-zusammenstellung	Testlänge	In der ICT-Skills Rahmenkonzeption definierte Facetten:				
		Situation			Soziale Interaktion	
		Persönlich	Beruflich	Bildungs-bezogen	Individuell	Kollektiv
Manuell	10	4	3	3	5	5
	15	5	4	6	8	7
	20	5	7	8	11	9
	25	5	9	11	14	11
Automatisiert	10	2	3	5	3	7
	15	5	4	6	6	9
	20	7	4	9	10	10
	25	7	7	11	14	11

Anmerkungen. Die fünf kognitiven Prozesse werden in jedem Test gleichverteilt abgedeckt.

Die größte Abweichung lässt sich für die 10 Items umfassenden Tests in Bezug auf die Verteilung von Items erkennen. Hier wurde bei der manuellen Testzusammenstellung noch stärker auf eine Gleichverteilung der Items für die Facettenebenen geachtet. Bei automatisierter Testzusammenstellung ergibt sich beispielsweise bei dem 10 Items umfassenden Kurztest eine stärkere Gewichtung von Items, die auf die Zusammenarbeit mit anderen abzielten (kollektiv), wobei aber die gesetzte Nebenbedingung eine stärkere Ausgewogenheit über die Facettenebenen auch nicht erforderte. Weiterhin wurde auf eine möglichst hohe Vielfalt gängiger Computeranwendungen bereits mit der ersten (bzw. kürzesten) Testzusammenstellung geachtet. So erstrecken sich die Umgebungen, mit denen in den Items aller ICT-Kurztests umzugehen ist, von Internetseiten im Web-Browser über E-Mail-Postfach und Ordner-Struktur bis hin zur Präsentationssoftware, dem Textverarbeitungs- oder Tabellenkalkulationsprogramm.

Diese Testzeiten wurden durch das Heranziehen der mittleren Bearbeitungszeit der jeweiligen ICT-Items aus der Kalibrierungsstudie geschätzt und werden in Tabelle 8.4 dargestellt.

Tabelle 8.4

Zusammenschau und Vergleich der verschieden ICT-Kurztests hinsichtlich der Testzeiten mit Darstellung der kleinsten (Min) und größten (Max) Itembearbeitungszeit sowie Mittelwert (M) und Standardabweichung (SD) nach Ansatz bei der Kurztestzusammenstellung und Testlänge

Kurztest-zusammenstellung	Test-länge	Testzeit (Min.)	Itembearbeitungszeit (Sek.)			
			Min	Max	*M*	*SD*
Manuell	10	17.70	62.63	149.49	106.19	34.73
	15	26.99	62.63	158.35	107.95	33.62
	20	34.69	58.59	158.35	104.07	32.77
	25	46.20	58.59	214.55	110.88	37.68
Automatisiert	10	19.76	54.10	214.55	118.58	43.97
	15	28.31	54.10	214.55	113.23	37.74
	20	36.93	54.10	214.55	110.80	36.91
	25	46.08	54.10	214.55	110.60	36.08

Die vier manuell zusammengestellten Tests veranschlagen zur Bearbeitung jeweils Zeit von circa 18, 27, 35 und 46 Minuten, die automatisiert zusammengestellten Tests von jeweils 20, 28, 37 und 46 Minuten. Die jeweiligen Testzeiten der manuell und automatisiert zusammengestellten ICT-Kurztests unterschieden sich demnach kaum und liegen fast alle deutlich unter den maximal definierten Testzeiten von 20, 30, 40 und 50 Minuten.

8.1.2 Studiendesign

Für Studie III ergibt sich ein $2 \times 4 \times 4$ Design, welches in Tabelle 8.5 dargestellt wird.

Tabelle 8.5

Studiendesign zum Vergleich der beiden unterschiedlichen Ansätze bei der Zusammenstellung von ICT-Kurztests verschiedener Länge unter Berücksichtigung von fehlenden Werten in den Antwortmatrizen

Kurztest-zusammenstellung	**Einbezug fehlender Werte**	**Testlänge**			
		10 Items	15 Items	20 Items	25 Items
Manuell	Keine Missings				
	OR				
	NR				
	OR & NR				
Automatisiert	Keine Missings				
	OR				
	NR				
	OR & NR				

Anmerkungen. OR = Omitted Responses; NR = Not Reached; OR & NR = Vorliegen von Omitted Responses und Not Reached.

Als unabhängige Variablen gelten dabei die Art der *Kurztestzusammenstellung*, die jeweilige *Testlänge* der ICT-Kurztests (mit 10, 15, 20 und 25 Items) und der *Einbezug fehlender Werte* (im Folgenden auch „Missing-Bedingung"). Zur Realisierung des Versuchsplans wurden zunächst manuell und computerbasiert automatisiert unterschiedlich lange Kurztests unter Berücksichtigung verschiedener Nebenbedingungen aus dem ICT-Itempool zusammengestellt (Kapitel 8.1.1). Anschließend wurden Personenparameter sowie Antworten auf die ausgewählten Items generiert und damit wiederum Personenfähigkeiten geschätzt. Bei der Generierung der Antwortmatrizen wurden auch fehlende Werte einbezogen, um differenzierte und belastbare Erkenntnisse aus den

Ergebnissen der Simulationsstudie ziehen zu können. Neben einer Bedingung mit vollständiger Antwortmatrix (keine Missings) wurden fehlende Werte entsprechend dem im CavE-ICT-Feldtest beobachteten Anteil ausgelassener Antworten (engl. Omitted Responses, OR), nicht erreichter und daher nicht beantworteter Items (engl. Not Reached, NR) sowie dem gemeinsamen Vorliegen von OR und NR berücksichtigt. Im Folgenden wird zunächst erläutert, wie fehlende Werte einbezogen wurden, bevor das Vorgehen der Datengenerierung erläutert wird.

Berücksichtigung von fehlenden Werten. Datensätze, die im Rahmen von psychologischen Untersuchungen erhoben werden, sind oft unvollständig (Lüdtke, Robitzsch, Trautwein & Köller, 2007). Um die Ergebnisse dieser Simulation zur Beurteilung der unterschiedlich langen ICT-Kurztests belastbarer und auch generalisierbar zu machen, sollen mögliche fehlende Werte im Datensatz über die Generierung von unvollständigen Antwortmatrizen einbezogen werden. In diesem Sinne wurde der Anteil fehlender Werte aus dem CavE-ICT-Feldtest herangezogen. Abbildung 8.3 gibt einen Überblick zum im Feldtest beobachteten prozentualen Anteil von fehlenden Werten für Personen (zu Missings und deren Kodierung im Rahmen des CavE-ICT-Feldtests siehe Kapitel 4.2.2). Werte, die aufgrund des verwendeten unvollständigen Testheftdesigns fehlen (Missings by design), liegen durch das im CavE-ICT-Feldtest verwendete Testheftdesign auch vor, werden in diese Betrachtungen allerdings nicht einbezogen, da diese Art der Missings im späteren Kurztest nicht auftreten kann. In Abbildung 8.3 ist daher die prozentuale Anzahl ausgelassener Antworten (OR) sowie nicht erreichter und daher nicht beantworteter Items (NR) abgebildet. Es ist festzuhalten, dass nur 16.06 % der Teilnehmenden weder Missings auf Grund von OR noch NR hatten und damit vollständige Datensätze lieferten. 42.82 % der am CavE-ICT-Feldtest teilnehmenden Schülerinnen und Schüler haben Aufgaben ausgelassen (Personen mit einem Anteil an OR-Missings im Datensatz der größer ist als Null), 71.28 % konnten nicht alle der ihnen

vorgelegten Aufgaben bearbeiten (Personen mit einem Anteil an NR-Missings im Datensatz der größer ist als Null).

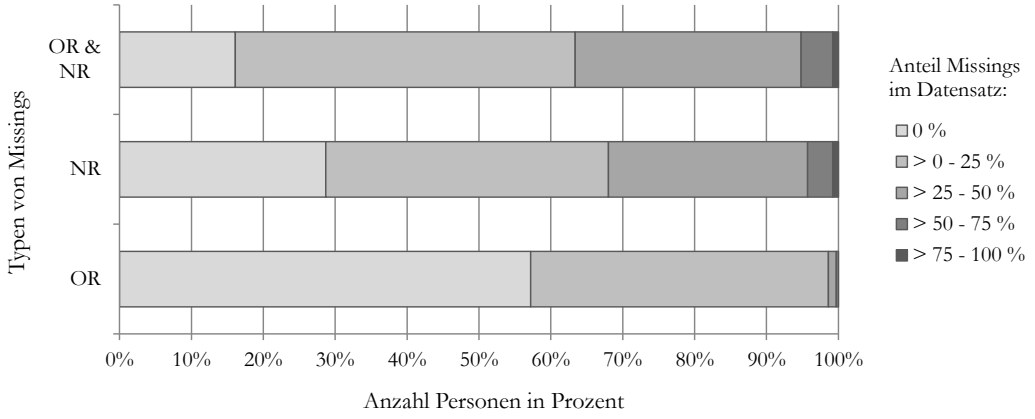

Abbildung 8.3 Anzahl von Personen in Prozent und deren jeweiliger Anteil von Missings (OR = Omitted Responses; NR = Not Reached; OR & NR = Vorliegen von Omitted Responses und Not Reached) in den Daten.

Betrachtet man den kompletten Datensatz der 766 am CavE-ICT-Feldtest teilnehmenden Schülerinnen und Schüler, zeigt sich insgesamt ein Anteil von 2.70 % OR, 18.14 % NR und 20.83 % fehlender Werte auf Grund von OR und NR in Bezug auf die 64 selektierten ICT-Items. Im Rahmen der Simulationsstudie werden daher neben einer vollständigen Antwortmatrix (keine fehlenden Werte) auch Antwortmatrizen mit unterschiedlichen Anteilen von fehlenden Werten erzeugt. Dabei soll jeweils der Anteil von OR und der Anteil von NR sowie die Kombination beider Typen von fehlenden Werten einbezogen werden. Da der Anteil beobachteter OR und unter Umständen auch NR (z.B. wenn durch langsames Arbeiten nicht alle Items in der zur Verfügung stehenden Testzeit bearbeitet werden konnten) – nicht per se zufällig über Personen auftreten, sondern in Zusammenhang mit der Personenfähigkeit stehen kann(Lüdtke et al., 2007), sollen die im CavE-ICT-Feldtest beobachteten Anteile fehlender Werte nicht zufällig über die Antwortmatrix verteilt werden. Um diesen Anspruch zu realisieren, wurde eine

Regression des Anteils fehlender Werte (OR und NR) auf die ICT-Fähigkeit durchgeführt. Es zeigt sich für den Anteil an OR $\beta = -.373$ $(t(765) = 14.703, p < .001)$, für den Anteil an NR $\beta = .082$ $(t(765) = 28.212, p = .023)$ und für die Kombination beider Missing-Typen $\beta = -.034$ $(t(765) = 32.531, p < .001)$. Für die Simulation werden die Ergebnisse der Regressionsanalysen genutzt, um aus der generierten Personenfähigkeit den jeweiligen Anteil fehlender Werte vorherzusagen. Die oben genannten Berechnungen implizieren folgende Regressionsgleichungen zur Vorhersage des Anteils fehlender Werte:

$$OR_{\text{predict}} = .027 - .031 \cdot \theta, \tag{8.17}$$

$$NR_{\text{predict}} = .179 + .022 \cdot \theta \tag{8.18}$$

und

$$OR \,\&\, NR_{\text{predict}} = .206 - .009 \cdot \theta. \tag{8.19}$$

In der zunächst als vollständig simulierten Antwortmatrix wurde dann die vorhergesagte Anzahl von Antworten für jede Person zufällig durch fehlende Werte ersetzt.

Datengenerierung. Die Generierung der Daten und die Schätzung der IRT-Personenparameter erfolgte mit der Software R (R Version 3.3.1; R Core Team, 2016) und dem R-Paket MIRT (MIRT Version 1.23; Chalmers, 2012). Zur Generierung der Personenparameter wurden für jede der Bedingungen Stichproben von jeweils 1000 Personen aus einer Normalverteilung mit einem Mittelwert von 0 und einer Standardabweichung von 0.57 gezogen. Diese Werte wurden verwendet, da die Varianz der Personenparameter im CavE-ICT-Feldtest bei 0.33 lag, wobei der Mittelwert der Personenparameter auf Null fixiert wurde (Wenzel et al., 2016). Im nächsten Schritt wurden Antworten aller Personen auf die für den jeweiligen Kurztest ausgewählten Items generiert. Dabei wurde ein eindimensionales Rasch-Modell verwendet. Für die weiteren Missing-Bedingungen wurden entsprechend des zuvor beschriebenen Vorgehens die

Antwortmatrizen mit fehlenden Werten versehen. Um der statistischen Unsicherheit der simulierten Antwortprozesse Rechnung zu tragen, wurden 100 Replikationen durchgeführt, wobei für jede Versuchsbedingung und Replikation eine neue Stichprobe gezogen wurde. Zur Ermittlung der geschätzten ICT-Fähigkeiten wurde das R-Paket TAM (TAM Version 2.8-21; Robitzsch, Kiefer & Wu, 2017) genutzt. Gegeben der jeweiligen Itemschwierigkeiten und generierten Antworten wurden unter Nutzung eines eindimensionalen Rasch-Modells Personenparameter als WLE-Schätzwerte ermittelt. WLE-Schätzer wurden MLE-Schätzern vorgezogen, weil diese bei kürzeren Testlängen weniger verzerrt sind und auch Werte für Personen liefern, die kein oder jedes Item richtig beantworteten (vgl. Kapitel 2.2.1).

8.1.3 Evaluationskriterien

Zum Vergleich der verschieden langen ICT-Kurztests wurden vier Evaluationskriterien als abhängige Variablen berechnet. Ermittelt wurden, wie auch in den Studien I und II dieser Arbeit (Kapitel 6 und 7), *Bias*, *MSE* und *Reliabilität* sowie als weiteres Kriterium die *Messeffizienz (ME)*. Da im Zuge der vorliegenden Studie allerdings die Erfassung des Gesamtkonstrukts ICT-Skills im Vordergrund steht, werden die kognitiven Prozesse nicht mehr differenziert betrachtet, wodurch sich die Formalisierung der Kriterien zum Teil vereinfacht. Es ergibt sich daher für den Bias, als Mittelwert der Differenz zwischen dem geschätzten und wahren Personenparameter:

$$Bias = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)}{N}. \tag{8.20}$$

Für den MSE ergibt sich:

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)^2, \tag{8.21}$$

wobei $\hat{\theta}_j$ der geschätzten und θ_j der wahren Fähigkeit für Person j entspricht.

Zusätzlich soll über die Betrachtung des MSE bedingt auf die Personenfähigkeiten auch etwas zur Güte der erzielten Schätzungen in allen Bereichen des Fähigkeitsspektrums ausgesagt werden. Als weiteres Kriterium zum Vergleich der verschiedenen ICT-Kurztests beziehungsweise Testlängen soll die ME herangezogen werden. Diese ist als Quotient von Messpräzision und Testlänge definiert (Frey & Ehmke, 2007), wobei die Messpräzision durch den Kehrwert des MSE ausgedrückt werden kann. Es ergibt sich entsprechend:

$$ME = \frac{MSE^{-1}}{m}, \tag{8.22}$$

wobei m die Testlänge wiedergibt. Höhere Werte sprechen für eine bessere Messeffizienz.

Die Reliabilität der Personenparameterschätzungen wurde wiederum als quadrierte Korrelation geschätzter und wahrer Personenparameter ermittelt:

$$REL = r^2_{\hat{\theta}_j,\theta_j}. \tag{8.23}$$

Während sich die Einschätzung der errechneten Reliabilitäten in den Studien I und II an den bei PISA berichteten Werten (min. .55) orientierte, werden in Studie III noch weitere Empfehlungen berücksichtigt. Diese beziehen sich auf die Interpretation der Reliabilität von Testverfahren, die für Untersuchungen kleinerer Gruppen, zu Screening-Zwecken oder im Bereich der Individualdiagnostik eingesetzt werden sollen. Sollte der Kurztest zur Individualdiagnostik eingesetzt werden, ist eine hohe Reliabilität unverzichtbar. Beim Einsatz zu Screening-Zwecken ist hingegen zu bedenken, dass in diesem Kontext in der Regel auf Basis eines möglichst kurzen Tests eine grobe Einschätzung der individuellen Merkmalsausprägung vorgenommen werden soll. Daher werden bei Testverfahren mit diesem Anwendungsziel meist nicht sehr hohe Reliabilitäten antizipiert (Schermelleh-Engel & Werner, 2012). Wenn der ICT-Kurztest zur Kollektivdiagnostik eingesetzt werden sollte, kann ebenfalls eine etwas niedrigere

Reliabilität als ausreichend erachtet werden. Als Orientierung zur Einschätzung der Reliabilität soll daher ein Wert von etwa .70 als ausreichend, von .80 als gut und über .90 als hoch betrachtet werden (Danner, 2015).

Die zur Evaluation genutzten Kriterien wurden je Replikation berechnet und anschließend über alle Replikationen gemittelt, wobei für die Berechnung der mittleren Reliabilitäten Fisher-Z-Transformationen durchgeführt wurden.

8.2 Ergebnisse

8.2.1 Systematischer Messfehler (Bias)

In allen Versuchsbedingungen konnten geringe Messfehler beobachtet werden. Die gemittelten Werte des Bias sowie deren Standardfehler werden in Tabelle 8.6 für alle Versuchsbedingungen dargestellt.

Tabelle 8.6

Über die Replikationen gemittelter Bias und Standardfehler (SE) der verschieden langen ICT-Kurztests für die verschiedenen Missing-Bedingungen und Ansätze der Testzusammenstellung

Kurztest-zusammenstellung	Missing-Bedingung	Testlänge			
		10 Items Bias (*SE*)	15 Items Bias (*SE*)	20 Items Bias (*SE*)	25 Items Bias (*SE*)
Manuell	Keine Missings	-.001 (.023)	-.000 (.018)	-.001 (.017)	-.001 (.015)
	OR	-.002 (.024)	-.001 (.019)	-.001 (.017)	-.001 (.015)
	NR	-.002 (.026)	-.001 (.020)	-.002 (.019)	-.002 (.016)
	OR & NR	-.003 (.026)	-.002 (.020)	-.002 (.018)	-.002 (.016)
Automatisiert	Keine Missings	-.005 (.023)	-.001 (.018)	.000 (.017)	-.000 (.015)
	OR	-.006 (.023)	-.001 (.018)	.000 (.017)	-.000 (.015)
	NR	-.010 (.023)	-.003 (.019)	-.001 (.018)	-.002 (.016)
	OR & NR	-.011 (.024)	-.005 (.019)	-.002 (.018)	-.002 (.016)

Anmerkungen. OR = Omitted Responses; NR = Not Reached; OR & NR = Vorliegen von Omitted Responses und Not Reached

Über die verschiedenen Missing-Bedingungen hinweg unterschieden sich die Werte des Bias kaum und auch die Unterschiede in den Werten für die beiden verschiedenen Herangehensweisen der Kurztestzusammenstellung waren ähnlich, mit leichten Vorteilen für die manuelle Zusammenstellung bei kürzeren Tests. Insgesamt können die Fähigkeitsschätzungen aber im Mittel als unverzerrt beurteilt werden.

8.2.2 Mittlere quadrierte Abweichung – Mean Squared Error (MSE)

Deutlicher als beim Bias waren Unterschiede zwischen den Versuchsbedingungen bei Betrachtung der Werte des MSE zu erkennen. In Tabelle 8.7 sind die über 100 Replikationen gemittelten Werte des MSE mit Standardfehlern für die einzelnen Versuchsbedingungen abgetragen.

Tabelle 8.7

Über die Replikationen gemittelter Mean Squared Error (MSE) und Standardfehler (SE) der verschieden langen ICT-Kurztests für die unterschiedlichen Missing-Bedingungen und Ansätze der Testzusammenstellung

Kurztest-zusammenstellung	Missing-Bedingung	Testlänge			
		10 Items *MSE* (*SE*)	15 Items *MSE* (*SE*)	20 Items *MSE* (*SE*)	25 Items *MSE* (*SE*)
Manuell	Keine Missings	.562 (.032)	.361 (.021)	.265 (.015)	.208 (.012)
	OR	.579 (.034)	.374 (.022)	.272 (.015)	.214 (.013)
	NR	.688 (.043)	.448 (.031)	.326 (.022)	.256 (.016)
	OR & NR	.714 (.043)	.465 (.028)	.337 (.021)	.265 (.015)
Automatisiert	Keine Missings	.482 (.026)	.330 (.019)	.249 (.013)	.203 (.011)
	OR	.496 (.027)	.341 (.020)	.257 (.013)	.209 (.011)
	NR	.579 (.034)	.404 (.026)	.306 (.018)	.250 (.014)
	OR & NR	.598 (.033)	.419 (.025)	.317 (.017)	.260 (.013)

Anmerkungen. OR = Omitted Responses; NR = Not Reached; OR & NR = Vorliegen von Omitted Responses und Not Reached

Hier zeigte sich, dass mit steigender Itemzahl im Kurztest der MSE sank. Des Weiteren stiegen die Werte des MSE mit zunehmender Berücksichtigung möglicher fehlender Werte an. Zudem zeigten sich über alle Bedingungen hinweg niedrigere Werte des MSE für die automatisiert zusammengestellten Kurztests, wobei dieser Unterschied

bei geringer Testlänge deutlicher ausgeprägt war. So zeigten sich in den Bedingungen mit Berücksichtigung von Missings durch OR und NR für manuell zusammengestellte Kurztests jeweils die höchsten MSE Werte zwischen .714 (10 Items Testlänge) und .265 (25 Items Testlänge). Lagen die Antwortmatrizen hingegen vollständig vor, so konnten MSEs zwischen .482 und .203 für automatisiert zusammengestellte Kurztests ermittelt werden. Durch den Ansatz der automatisierten Testzusammenstellung konnte der MSE bei einer Testlänge von 10 Items um bis zu 16.2 % und bei einer Testlänge von 15 Items um bis zu 9.9 % im Vergleich zur manuellen Testzusammenstellung verringert werden (bei Einbezug von OR und NR). Bei längeren Kurztests fiel die Verringerung des MSE durch automatisierte im Vergleich zu manueller Testzusammenstellung mit Werten zwischen 6.1 % und 1.9 % etwas geringer aus.

Es zeigt sich, dass in allen Versuchsbedingungen für Personen an den Rändern der Fähigkeitsverteilung die Messfehler deutlich größer ausfielen als für Personen im mittleren Fähigkeitsbereich (Abbildung 8.4). Mit zunehmender Testlänge wurden die individuellen Messfehler generell kleiner. Ebenso konnten geringere Messfehler beobachtet werden, wenn die Antwortmatrizen weniger fehlende Werte aufwiesen. Aus Abbildung 8.4 kann zudem entnommen werden, dass vor allem der Einbezug des Anteils fehlender Werte durch NR zu einem Anstieg der Messfehler führte. Im Vergleich der beiden Ansätze zur Testzusammenstellung wurde deutlich, dass bei geringer Testlänge (10 und 15 Items) die Messfehler für die automatisiert zusammengestellten Tests über das gesamte Fähigkeitsspektrum hinweg etwas kleiner ausfielen. Bei den etwas längeren Tests (20 und 25 Items) zeigten sich im negativen Fähigkeitsbereich etwas geringere Messfehler für die manuell zusammengestellten Tests, wobei die automatisiert zusammengestellten Tests im mittleren Fähigkeitsbereich auch bei diesen Testlängen die kleinsten MSEs aufwiesen.

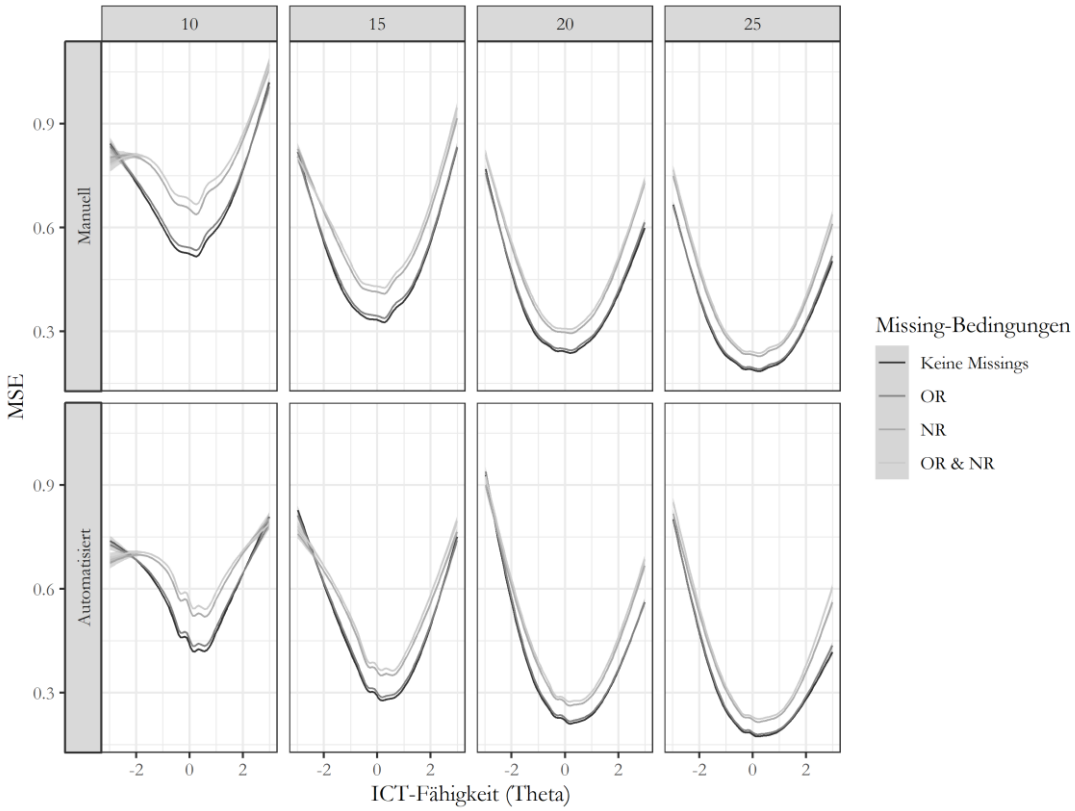

Abbildung 8.4 Mean Squared Error (MSE) bedingt auf die Personenfähigkeit in den vier Missing-Bedingung (OR = Omitted Responses; NR = Not Reached; OR & NR = Vorliegen von Omitted Responses und Not Reached) für die verschiedenen ICT-Kurztests (bestehend aus 10, 15, 20 oder 25 Items) und die unterschiedlichen Ansätze der Testzusammenstellung.

8.2.3 Messeffizienz (ME)

Tabelle 8.8 zeigt die über 100 Replikationen gemittelten Werte der ME mit Standardfehlern für die einzelnen Versuchsbedingungen.

Es wird deutlich, dass bei automatisierter Testzusammenstellung für kürzere Tests leicht bessere Werte der ME erzielt werden konnten als für längere ICT-Kurztests. Für manuell zusammengestellte Tests zeigt sich ein gegenläufiges Bild, hier steigt die ME mit der Testlänge leicht an.

Tabelle 8.8

Über die Replikationen gemittelte Messeffizienz (ME) und Standardfehler (SE) der verschieden langen ICT-Kurztests für die unterschiedlichen Missing-Bedingungen und Ansätze der Testzusammenstellung

Kurztest-zusammenstellung	Missing-Bedingung	Testlänge: 10 Items *ME* (*SE*)	15 Items *ME* (*SE*)	20 Items *ME* (*SE*)	25 Items *ME* (*SE*)
Manuell	Keine Missings	.179 (.010)	.185 (.011)	.190 (.011)	.193 (.011)
	OR	.173 (.010)	.179 (.011)	.184 (.010)	.188 (.011)
	NR	.146 (.009)	.149 (.010)	.154 (.010)	.157 (.010)
	OR & NR	.141 (.009)	.144 (.009)	.149 (.009)	.151 (.009)
Automatisiert	Keine Missings	.208 (.011)	.203 (.012)	.201 (.010)	.197 (.010)
	OR	.202 (.011)	.196 (.011)	.195 (.010)	.192 (.010)
	NR	.173 (.010)	.166 (.011)	.164 (.010)	.160 (.009)
	OR & NR	.168 (.009)	.160 (.010)	.158 (.009)	.155 (.008)

Anmerkungen. OR = Omitted Responses; NR = Not Reached; OR & NR = Vorliegen von Omitted Responses und Not Reached

Unterschiede im Vergleich der ME zwischen den beiden Ansätzen der Testzusammenstellung zeigten sich für alle ICT-Kurztest zugunsten der automatisiert erstellten Tests. Diese können lediglich im Vergleich der Tests mit 25 Items als nicht wesentlich bezeichnet werden, beurteilt anhand der jeweiligen Standardfehler. Die Verbesserung der ME bei automatisierter im Vergleich zu manueller Testzusammenstellung lag bei bis zu 19.1 %. Bei 15 Items Testlänge konnte eine Steigerung der ME um bis zu 11.4 % beobachtet werden, bei 20 Items Testlänge noch um bis zu 6.5 % und bei 25 Items Testlänge bei maximal 2.6 %. Bei Betrachtung der verschiedenen Missing-Bedingungen schnitten die Bedingungen, in denen NR und die Kombination aus NR und OR berücksichtigt wurden, deutlich schlechter ab als jene ohne fehlende Werte oder mit einem vergleichsweise geringeren Anteil fehlender Werte durch OR.

8.2.4 Reliabilität

Tabelle 8.9 stellt die über Replikationen gemittelten Reliabilitäten der ICT-Kurztests in den verschiedenen Versuchsbedingungen mit zugehörigen Standardfehlern dar. In

Hinblick auf die Reliabilitäten der verschiedenen ICT-Kurztests in den Versuchsbedingungen ließ sich erkennen, dass die höchsten Reliabilitäten bei einem geringen Missinganteil ermittelt wurden.

Tabelle 8.9
Über die Replikationen gemittelte Reliabilität (REL) und Standardfehler (SE) der verschieden langen ICT-Kurztests für die unterschiedlichen Missing-Bedingungen und Ansätze der Testzusammenstellung

Kurztest-zusammenstellung	Missing-Bedingung	Testlänge: 10 Items *REL* (*SE*)	15 Items *REL* (*SE*)	20 Items *REL* (*SE*)	25 Items *REL* (*SE*)
Manuell	Keine Missings	.503 (.027)	.613 (.024)	.684 (.025)	.734 (.027)
	OR	.495 (.027)	.610 (.024)	.678 (.024)	.728 (.027)
	NR	.450 (.028)	.561 (.027)	.637 (.028)	.691 (.029)
	OR & NR	.440 (.027)	.551 (.024)	.629 (.025)	.683 (.027)
Automatisiert	Keine Missings	.535 (.029)	.632 (.028)	.696 (.027)	.737 (.029)
	OR	.527 (.030)	.625 (.028)	.689 (.027)	.731 (.029)
	NR	.484 (.032)	.582 (.031)	.649 (.031)	.694 (.031)
	OR & NR	.474 (.031)	.573 (.030)	.641 (.029)	.687 (.029)

Anmerkungen. OR = Omitted Responses; NR = Not Reached; OR & NR = Vorliegen von Omitted Responses und Not Reached

Erwartungsgemäß steigt die Reliabilität mit der Testlänge an. Unterschiede zwischen den beiden Ansätzen der Testzusammenstellung zeigten sich kaum. Der erzielte Zugewinn durch die Nutzung automatisiert zusammengestellter Tests im Vergleich zu manuell zusammengestellte nahm mit der Testlänge ab. Bei den 10 Items umfassenden Kurztests lag er noch bei bis zu 7.7 %, bei 15 Items Testlänge bei maximal 4.0 % und bei 20 und 25 Items nur noch bei bis zu 1.9 % beziehungsweise 0.6 %. Generell kann festgehalten werden, dass mit steigendem Missinganteil in den Antwortmatrizen die Reliabilitätswerte jeweils abnahmen. Für die Bedingungen ohne Missings wurden dementsprechend die höchsten Reliabilitäten ermittelt, mit Werten zwischen .503 und .734 für manuell beziehungsweise zwischen .535 und .737 für automatisiert zusammengestellte Kurztests.

8.3 Diskussion

Ziel der dritten Studie der vorliegenden Arbeit war es, eine Empfehlung für einen Kurztest abzuleiten, mit dem zum einen die in Bezug auf die aufgewendete Testzeit ökonomische aber zum anderen auch die messeffiziente, präzise und reliable Erfassung von ICT-Skills ermöglicht wird. Über die gezielte multikriteriale Testzusammenstellung wurde versucht, dem Anspruch möglichst valider Testwertinterpretation gerecht zu werden. Dabei wurden zwei unterschiedliche Ansätze der Testzusammenstellung gegenübergestellt, zum einen ein manuelles und zum anderen ein computerbasiertes automatisiertes Vorgehen. Zunächst soll kurz das unterschiedliche Vorgehen bei der Testzusammenstellung bewertet werden. Die manuelle Testzusammenstellung ist durchaus zeitaufwändig und eher unflexibel beispielsweise im Hinblick auf das Überdenken und Anpassen von Nebenbedingungen. Der in dieser Studie genutzte CavE-ICT-Itempool ist mit seinen 64 Item vergleichsweise überschaubar groß. Bei üblicherweise größeren Itempools ist davon auszugehen, dass es noch komplexer ist, genügend Überblick über die verschiedenen Items und ihre Eigenschaften zu haben, um diese möglichst fundiert auszuwählen. Was sich allerdings bereits im Rahmen dieser Studie zeigte, ist dass die große Anzahl an Nebenbedingungen schon dazu führt, dass die händische Itemselektion sehr schwierig, zeitaufwändig und fehleranfällig ist. Eine computerbasierte automatisierte Testzusammenstellung hat unter anderem den Vorteil, dass Nebenbedingungen, sobald sie formalisiert sind, auch leichter angepasst und Effekte der Veränderung von Nebenbedingungen untersucht werden können. Das bietet die Möglichkeit, Vergleiche zwischen der Setzung restriktiverer oder liberalerer Nebenbedingungen vorzunehmen. Ähnliches gilt für die Definition der Zielsetzung in Form einer anvisierten TIF. Diese kann ebenfalls ohne großen Aufwand angepasst werden, beispielsweise um den Fokus der Messgenauigkeiten zwischen den Extremwerten oder der Mitte der Fähigkeitsverteilung zu variieren. Dies ist vor allem attraktiv, da die

eigentliche Zeit der Testzusammenstellung (also die Rechenzeit) sehr kurz ausfällt. Zeitintensiv ist vor allem die vorgelagerte inhaltliche Überlegung zur Setzung von Nebenbedingungen und Zielen, wie sie allerdings auch für die manuelle Testzusammenstellung erforderlich ist. Auf der anderen Seite ist für die Nutzung der Vorteile der computerbasiert automatisierten Testzusammenstellung auch der Zugang beziehungsweise das Verständnis und die Fähigkeit zu entsprechenden Programmierungen erforderlich. Diese Hürde ist sicherlich als so hoch zu bewerten, dass die händische Zusammenstellung von Kurztests als praktikable Lösung oft gewählt wird, vor allem wenn beispielsweise die Einhaltung von Nebenbedingungen bei der Testzusammenstellung keine so große Rolle spielt. Darüber hinaus muss auch festgehalten werden, dass der im Rahmen dieser Studie manuell zusammengestellte Kurztest weniger stark als klassische Ansätze der Itemauswahl zur Erstellung von Skalen (vgl. Eid und Schmidt, 2014) auf spezifische Eigenschaften der Items fokussierte, sondern zugleich auch das Zusammenwirken der ausgewählten ICT-Items als Kurztest in den Blick nahm. Dieser Fokus bei der Itemauswahl ist in der hier realisierten automatisierten Formen der Testzusammenstellung aber nochmal deutlich stärker gesetzt. Allerdings ist die im Rahmen dieser Studie dargestellte und angewendete Form der automatisierten Testzusammenstellung als lineares Optimierungsproblem nach van der Linden (2005) nur eine mögliche Variante. Beispielsweise kann auch der stuart-Ansatz zur Zusammenstellung von Kurztests verfolgt werden, bei dem ein Ant-Colony-Algorithmus zur optimalen Testzusammenstellung genutzt wird (Schultze, 2017). Um die vier manuell und die vier automatisiert zusammengestellten ICT-Kurztests zu evaluieren, wurden anschließend an die Testzusammenstellung Simulationen durchgeführt, wobei durch den Einbezug möglicher fehlender Werte die Generalisierbarkeit und Belastbarkeit der Ergebnisse weiter erhöht wurde. Es zeigt sich über alle betrachten Evaluationskriterien hinweg, dass die automatisiert erstellten Tests etwas besser abschneiden als die manuellen Kurztests. Es

konnte weiterhin gezeigt werden, dass mit zunehmender Itemzahl der Kurztests die Messfehler geringer wurden und zum anderen die Reliabilitäten stiegen. Eine generell präzisere Messung durch längere ICT-Kurztests lässt sich über die Betrachtung der MSEs feststellen, wobei sich auch die Messeffizienz der verschiedenen Tests unterscheidet. Die individuellen Messfehler für einzelne Personen sinken bei den längeren Tests deutlich und sind insgesamt etwas homogener verteilt. Festzuhalten ist dabei, dass für jeden der vergleichend betrachteten Tests die Schätzfehler in den Randbereichen der Verteilung deutlich von denen im mittleren Bereich des Fähigkeitsspektrums verschieden sind. Die Berücksichtigung verschiedener Anteile von fehlenden Werten bei den Simulationen zeigt den nicht zu vernachlässigen Effekt unvollständiger Daten auf die Testgüte und sollte einer Überschätzung dieser entgegenwirken. Um eine Empfehlung für einen der insgesamt acht zusammengestellten ICT-Kurztests abzuleiten, ist eine Abwägung zwischen den erzielten Ergebnissen zur Testevaluation, der antizipierten durchschnittlichen Testzeit und des Einsatzzwecks des Tests vorzunehmen. Für individualdiagnostische Zwecke akzeptable Reliabilitätswerte von über .70 konnten bei einer angenommen Fähigkeitsverteilung wie in der CavE-ICT-Feldtest-Stichprobe nur für die 25 Items umfassenden Kurztests ermittelt werden und dies auch nur für die Versuchsbedingungen mit keinem oder einem geringen Anteil an Missings. Generell sollte daher überlegt werden, inwieweit sich Missings möglichst minimieren lassen. Im Rahmen des CavE-ICT-Feldtests wurden relativ wenige OR abgegeben. Dies liegt wahrscheinlich auch daran, dass das Überspringen eines Items nicht mit einem Klick möglich war, sondern vor dem Weitergehen zum nächsten Item dies nochmal explizit bestätigt werden musste. Ein unbeabsichtigtes Weiterklicken konnte somit wahrscheinlich meist verhindert werden. Der verhältnismäßig hohe Anteil von NR in den CavE-ICT-Feldtestdaten könnte dafür sprechen, dass die Testzeit nicht ausreichend war, um alle Items zu bearbeiten. Tatsächlich standen 50 Minuten Testzeit zur Verfügung, in denen die Schülerinnen und Schüler

zwischen 29 und 35 ICT-Items vorgelegt bekamen, dies würde im Schnitt eine Bearbeitungszeit von 93.75 Sekunden pro Item erlauben. Tatsächlich lag die durchschnittliche Bearbeitungszeit der 64 insgesamt administrierten Items aber bei 105.60 Sekunden (vgl. Kapitel 4.3.1). Für den späteren Einsatz des ICT-Kurztests kann dieser zusätzlichen Speedkomponente während der Testung entgegengewirkt werden. Zum einen ist die Abschätzung der Testzeit durch Verwendung der im Feldtest beobachteten Itembearbeitungszeiten genauer, zum anderen könnte die Testzeit zusätzlich um einige Minuten verlängert werden, um so den Anteil von NR deutlich zu senken. Dennoch bleibt festzuhalten, dass auch mit Bemühungen, den Missinganteil im später einzusetzenden ICT-Kurztest möglichst gering zu halten, eine hohe Reliabilität von über .80 entsprechend der vorgestellten Ergebnisse kaum erreicht werden kann. Dies, wie auch die weniger genaue Schätzung von Personenparametern an den Rändern der Fähigkeitsverteilung, spricht eher gegen den Einsatz des ICT-Kurztests zu individualdiagnostischen Zwecken, für die eine hohe Messgenauigkeit für Personen jeder Fähigkeitsausprägung unverzichtbar ist. Für Screening-Zwecke, bei denen zugunsten eines zeitlich geringeren Aufwands eine niedrige Reliabilität in Kauf genommen werden kann, kann einer der vorgeschlagenen ICT-Kurztests angewendet werden. Auch für Gruppenvergleiche ist das Instrument geeignet, da die Gruppenmittelwerte auch bei individuell stärker messfehlerbehafteten Testwerten korrekt geschätzt werden (Schermelleh-Engel & Werner, 2012). Darüber hinaus wäre auf Basis der erzielten Ergebnisse dieser Studie auch im Kontext von LSAs der Einsatz eines ICT-Kurztests möglich. Bei dem hier dargestellten Vergleich der acht ICT-Kurztests anhand von statistischen Evaluationskriterien, deskriptiven Testeigenschaften (vornehmlich der unterschiedlichen zu veranschlagenden Testzeit) sowie Überlegungen zu Einsatzbedingungen und möglichen -bereichen des Instruments scheint der aus 20 Items automatisiert zusammengestellte Kurztest am praktikabelsten zu sein. Die Werte des Bias und des MSE fallen verhältnismäßig niedrig aus. Sofern zudem

der Anteil von fehlenden Werten reduziert werden kann, erreicht dieser ICT-Test eine Reliabilität von .689 (für die OR Missing-Bedingung). Mit einer antizipierten Testzeit von circa 37 Minuten ist er zudem gut als Kurztest einsetzbar. Die Reliabilitätssteigerung bei der Itemauswahl des Kurztests mit einer Testlänge von 25 ist zwar deutlich (.731 in der OR Missing-Bedingung), andererseits veranschlagt dieser Test jedoch circa 46 Minuten und würde inklusive des in die Testumgebung einführenden und erklärenden Trainings mindestens 55 Minuten Durchführungszeit veranschlagen. Dies spricht für die Auswahl des 20 Items umfassenden Tests, der mit Tutorial knapp innerhalb einer 45-minütigen Schulstunde durchgeführt werden könnte.

Bei einer späteren Anwendung des Testverfahrens sollten allerdings verschiedene weitere Aspekte bedacht werden, die im Zuge der vorliegenden Studie nicht näher untersucht wurden. Über die Zeit können Änderungen in den Itemparametern aber auch im latenten Merkmal ICT-Skills selbst auftreten und zu ungenauen Testwerten beziehungsweise Testwertvergleichen führen. Auch wenn das angewandte Rasch-Modell recht stabil gegenüber solchen Veränderungen ist (Babcock & Albano, 2012), wird in den *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014) empfohlen, Skalen regelmäßig auf ihre Stabilität hin zu überprüfen. Gerade im Hinblick auf das gegeben der rasanten digitalen Entwicklungen des 21. Jahrhunderts eher schnell veränderliche Konstrukt der ICT-Skills scheint diese Empfehlung berechtigt. Des Weiteren wurden im Zuge dieser Studie keine Itempositions- oder -reihenfolgeeffekte untersucht. Im Gegensatz zum CavE-ICT-Feldtest, in der über den Einsatz eines balancierten Testheftdesigns versucht wurde diese Effekte zu kontrollieren, könnten solche bei einem linear administrierten ICT-Kurztest zum Tragen kommen. Beispielsweise kann die Lösungswahrscheinlichkeit eines Items geringer (z.B. durch Ermüdung) oder höher (z.B. durch Übungseffekte) ausfallen, wenn dieses näher

zum Testende hin positioniert ist (Davis & Ferdous, 2005; Schweizer, Schreiner & Gold, 2009). Um derartige Störquellen zu vermeiden, die auf die Messung von ICT-Skills einwirken können, könnte eine randomisierte Itemvorgabe sinnvoll und durch die computerisierte Testung auch leicht umsetzbar sein. Dies ist aber nur dann sinnvoll, wenn mit den Testergebnissen keine Aussagen auf Individualebene getroffen werden sollen. Ist dies doch gewünscht, sollte die Reihenfolge der Itemvorgabe im Test zwischen Personen stabil gehalten werden, um vergleichbare Testwertinterpretationen zu ermöglichen.

Schließlich wurde über die gezielte Itemauswahl zur Zusammenstellung des ICT-Kurztest versucht, valide Testwertinterpretation sicherzustellen. Für den 64 Items umfassenden ICT-Gesamttest wurden Validierungsuntersuchungen vorgenommen, die weitere zu Validierungszwecken im Rahmen des CavE-ICT-Feldtests erhobene Instrumente nutzen (Engelhardt et al., 2019). Ähnliche Analysen könnten auch nur für die ausgewählten Items vorgenommen und die Ergebnisse mit denen für den Gesamttest verglichen werden.

9 Allgemeine Diskussion

Im Rahmen dieser Arbeit wurde untersucht wie ein komplexes Konstrukt, wie das der ICT-Skills, auf Basis der Item-Response-Theorie und unter Einsatz computerisierter Messinstrumente erfasst werden kann. Neben einer zuverlässigen und effizienten Messung galt es zudem, der Konstruktrepräsentation von ICT-Skills gerecht zu werden. Dabei setzten die empirischen Studien dieser Arbeit unterschiedliche Testformen um und sie knüpften an unterschiedlichen Punkten im Prozess der Testentwicklung an. In den ersten beiden Studien wurde der Frage nachgegangen, wie adaptive Testalgorithmen zur Erfassung von ICT-Skills eingesetzt werden können. Die dritte Studie beschäftigte mit der Zusammenstellung eines verkürzten linearen ICT-Skills-Messinstruments.

Studie I setzte noch vor der Entwicklung von Items zur Messung von ICT-Skills an und zielte darauf ab, Hinweise zum Umfang des zu erstellenden ICT-Itempools und zur Testlänge eines adaptiven Messinstruments bereitzustellen. Studie II baute direkt auf Studie I auf und nutzte die im Rahmen des Projekts CavE-ICT entwickelten und kalibrierten Items, beziehungsweise ihre ermittelten Itemeigenschaften zur weiteren Erprobung verschiedener CAT-Algorithmen. Dabei wurden die Facettenebenen der kognitiven Prozesse entsprechend des CavE-ICT-Frameworks über multidimensionales adaptives Testen berücksichtigt. Es wurden Möglichkeiten aufgezeigt, wie MAT zur Messung von ICT-Skills gewinnbringend eingesetzt werden kann und eine differenzierte Messung auf Ebene der verschiedenen kognitiven Prozesse von ICT-Skills erlaubt. Hierbei wurden explizit auch Möglichkeiten untersucht die Items sequentiell geordnet, entsprechend des durch sie repräsentierten kognitiven Prozesses von ICT-Skills, doch trotzdem adaptiv vorzugeben. Die durch Studie II erarbeiteten Erkenntnisse können insbesondere für die Erfassung von mehrdimensionalen Konstrukten oder facettierten Merkmalen in LSAs genutzt werden. Durch den Vergleich der Ergebnisse von Studie I, der simulierte Daten zugrunde lagen, und Studie II, in der mit Echtdaten aus dem

CavE-ICT-Feldtest gearbeitet wurde, ergeben sich zudem Implikationen für ein angemessenes Design von Simulationsstudien, die insbesondere noch vor der eigentlichen Itementwicklung ansetzen.

In Studie III wurden lineare Kurztests zur Messung von ICT-Skills zusammengestellt. Durch die gezielte Auswahl geeigneter, im CavE-ICT-Feldtest kalibrierter ICT-Items sollte bei möglichst geringer Testzeit zugleich hohe Messgenauigkeit und -zuverlässigkeit erreicht werden. Die in Studie III manuell und automatisiert computerbasiert zusammengestellten Tests wurden hinsichtlich eines Einsatzes sowohl auf Populationsebene, im Sinne einschlägiger LSAs, als auch darüber hinaus für individual- und gruppendiagnostische Zwecke evaluiert und Empfehlungen für den Kurztesteinsatz abgeleitet.

Im Folgenden werden die Ergebnisse der Studien integriert und anschließend zentrale Erkenntnisse dieser Arbeit kritisch reflektiert, bevor ein Fazit gezogen werden soll.

9.1 Integration der Studien

9.1.1 Konstruktannahmen und Datengrundlage

Die drei empirischen Studien dieser Arbeit nutzen Annahmen oder Daten, die im Rahmen des Projekts CavE-ICT erarbeitet wurden. Das im Projekt entwickelte CavE-ICT-Framework, genauer die Definition der internen Struktur von ICT-Skills als Konstrukt, wird in jeder Studie beziehungsweise in jeder der untersuchten Testformen zugrunde gelegt. In Studie I, die mit Beginn der Itementwicklung im CavE-ICT-Projekt konzipiert und durchgeführt wurde, wurden die fünf kognitiven Prozesse, welche bei der Lösung von ICT-Aufgaben relevant sein können, noch als psychometrisch trennbare Dimensionen angenommen. Ihre Messung mithilfe eines multidimensionalen adaptiven Tests sollte noch exploriert werden. Studie II setzt nach der Itementwicklung und -kalibrierung im Projekt an und evaluiert verschiedene MAT-Algorithmen unter

Nutzung der im Projekt erhobenen Daten und geschätzten Itemschwierigkeiten. Nach dem CavE-ICT-Feldtest bleibt allerdings festzuhalten, dass die fünf kognitiven Prozesse sich nicht als psychometrisch trennbare Dimensionen bestätigten, da das eindimensionale Rasch-Modell zur Schätzung der Itemschwierigkeiten angemessener erschien als ein fünfdimensionales Modell (siehe Kapitel 4). Dennoch wurden, aus inhaltlichen Überlegungen heraus, im Zuge der Feldtestanalysen, ICT-Skills zum einen als Gesamtwert für Personen, aber auch die ICT-Fähigkeit von Personen differenziert für die fünf kognitiven Prozesse berichtet. Um letzteres zu ermöglichen, wurden die eindimensional geschätzten Itemschwierigkeiten fixiert und ein fünfdimensionales Modell zur Schätzung der Personenfähigkeiten darauf angewendet. Untersuchungen hinsichtlich der Angemessenheit dieses Vorgehens, der jeweils gewählten Modelle oder die Anwendung anderer Modelle im Vergleich, wurden im Rahmen dieser Arbeit nicht unternommen, im folgenden Abschnitt werden dazu aber kritische Überlegungen angestellt. Für die Studien im Rahmen dieser Arbeit wurden die erzielten und berichteten Ergebnisse des CavE-ICT-Feldtests wie in Wenzel et al. (2015) publiziert als Grundlage verwendet. In Studie II steht allein die Rückmeldung von fünf Merkmalsausprägungen entsprechend der fünf kognitiven Prozesse im Fokus. Ein Gesamtwert für ICT-Skills wird nicht angegeben, das bedeutet, es wurde von einer Mittelwertbildung der fünf erzielten Schätzwerte abgesehen, da die jeweiligen Evaluationskriterien für die einzelnen Facetten mitunter deutlich unterschiedlich ausfielen. Allerdings bietet der eingesetzte CU-MAT eine eindimensionale Schätzung für ICT-Skills zusätzlich zur berichteten fünfdimensionalen Merkmalsschätzung. Somit könnte dieser Algorithmus der gewählten Berichtsmetrik im Rahmen des CavE-ICT-Feldtests entsprechen.

Durch Studie I und II wird wie beschrieben auf die kognitiven Prozesse fokussiert, allerdings wird nicht die gesamte Facettierung des ICT-Skills-Konstrukts damit abgebildet. Die Facetten „soziale Interaktion“ „Situation“ werden nicht in besonderer Weise in den

Tests beziehungsweise durch die Itemauswahl im Rahmen der adaptiven Algorithmen berücksichtigt. Wie viele Items der jeweiligen Facettenebenen schließlich in den adaptiven Tests administriert werden, wurde weder kontrolliert noch evaluiert. Dies liegt darin begründet, dass beide Facetten eher der inhaltlichen Strukturierung des Konstrukts dienen. Eine noch stärkere Einschränkung der adaptiven Itemauswahl durch weitere Content-Constraints ist theoretisch möglich und auch als sinnvoll anzusehen, um der komplexen Struktur des ICT-Skills Konstrukts gerecht zu werden, allerdings auf Basis des eher kleinen vorliegenden ICT-Itempools praktisch nicht sinnvoll realisierbar.

Im Zuge von Studie III hingegen wird die interne Struktur von ICT-Skills durch die zusammengestellten ICT-Kurztests umfänglich abgebildet. Dies wurde bei der Zusammenstellung explizit angestrebt. Durch den resultierenden Kurztest wird letztlich eine Gesamtfähigkeit geschätzt und keine Differenzierung nach kognitiven Prozessen vorgenommen.

9.1.2 Anwendungsbereich und Testlänge der jeweils empfohlenen Instrumente

Im Zuge des Projekts CavE-ICT sollte ein Test bereitgestellt werden, der vornehmlich im Rahmen von LSAs zur Populationsbeschreibung eingesetzt werden könnte. Auf Basis der CavE-ICT-Projektergebnisse lässt sich ableiten, dass der 64 ICT-Items umfassende Itempool bei linearer Testung eine psychometrisch abgesicherte und theoriekonforme Erfassung von ICT-Skills ermöglicht (Wenzel et al., 2016). Der Gesamttest erreicht eine Reliabilität von .796. Die Reliabilitäten für die Fähigkeitsschätzungen der verschiedenen kognitiven Prozesse liegen zwischen .639 und .727 (siehe Kapitel 4). Dabei müsste für diesen ICT-Test allerdings insgesamt eine Testzeit von etwa 113 Minuten veranschlagt werden. Zudem wurde im Rahmen des Projekts untersucht, ob ein eindimensionaler adaptiver Test eingesetzt werden könnte. Bei einer Gleichverteilung von Items entsprechend der kognitiven Prozesse ergibt sich bereits bei

einer CAT-Länge von 17 Items eine Reliabilität .550 bei einer geschätzten Testzeit von circa 30 Minuten; eine Reliabilität von .700 kann nach Vorgabe von 33 ICT-Items erreicht werden, die Testzeit beträgt dann schätzungsweise 53 Minuten. Mithilfe eines solchen eindimensionalen CAT könnte allerdings nur eine ICT-Gesamtfähigkeit rückgemeldet werden, eine Rückmeldung für die kognitiven Prozesse ist nicht möglich.

Im Rahmen von Studie I und II wurde der Fokus auf die Exploration von Möglichkeiten solcher, für die kognitiven Prozesse differenzierenden und populationsbeschreibenden Rückmeldungen gesetzt. Die Ergebnisse von Studie II sprechen für den Einsatz eines mindestens 40 Items umfassenden MAT – unabhängig davon, ob Items sequentiell nach Dimensionen oder gemischt administriert werden. Die Reliabilitäten der Fähigkeitsschätzungen für die kognitiven Prozesse liegen zwischen .491 und .663 (für die unterschiedlichen MAT-Algorithmen). Die für einen dieser adaptiven Tests erforderliche Testzeit läge bei schätzungsweise 70 Minuten.

In Studie III wurde dann die Ableitung eines ICT-Kurztests angestrebt, der die interne Struktur von ICT-Skills umfänglich abbildet, die Rückmeldung einer ICT-Gesamtfähigkeit ermöglicht und neben dem Einsatz zur Populationsbeschreibung im Rahmen von LSAs auch für Individual- und Gruppendiagnostik sehr ökonomisch einsetzbar wäre. Der im Resultat empfohlene Test setzt sich aus 20 ICT-Items zusammen und erreicht eine Reliabilität von .689 bei einer Testzeit von etwa 37 Minuten.

In den Studien II und III wurden unterschiedliche Testverfahren zur Messung von ICT-Skills unter Nutzung der im Rahmen des CavE-ICT-Feldtests geschätzten Itemschwierigkeiten und Personenparameterverteilung erprobt. Die Testverfahren zielten dabei wie bereits beschrieben auf unterschiedliche Rückmeldungen ab – eine ICT-Gesamtfähigkeit vs. differenzierte Rückmeldung von Fähigkeiten hinsichtlich der fünf kognitiven Prozesse – sie fokussierten den Einsatz für unterschiedliche Zwecke und bilden das ICT-Framework unterschiedlich detailliert ab. Es wird deutlich, dass je nach

Anwendungszweck und zugrundeliegender Berichtsmetrik die Testzeiten deutlich unterschiedlich ausfallen.

9.2 Kritische Reflexion

Im Folgenden werden die zentralen Ergebnisse dieser Arbeit kritisch reflektiert. Dazu wird zunächst generell auf die Erfassung von ICT-Skills und speziell auf die im Rahmen dieser Arbeit vorgeschlagenen Instrumente (MAT und ICT-Kurztest) eingegangen. Im Anschluss wird die Nützlichkeit von Simulationsstudien im Zuge der Testentwicklung reflektiert. Letztlich wird auf die Möglichkeiten und Grenzen verschiedener Ansätze von CAT eingegangen, welche die Multidimensionalität von zu messenden Konstrukten auf unterschiedliche Weise berücksichtigen.

9.2.1 Erfassung von ICT-Skills

Da die Nutzung von ICT im Alltag allgegenwärtig ist und die Bedeutsamkeit von Fähigkeiten im Umgang mit ICT für die Teilhabe an der modernen Wissensgesellschaft außer Frage steht, ist es wichtig, ICT-Skills auch valide erfassen zu können, um beispielsweise Disparitäten und Interventionsbedarfe früh zu erkennen und in alltäglichen Bildungsprozessen aktiv gegensteuern zu können.

Die alleinige Nutzungshäufigkeit von ICT stellt nicht per se einen kompetenten Umgang sicher. Auch die Generation der „Digital Natives“ ist nicht automatisch auch kompetent im Umgang mit ICT. Mit Blick auf den wichtigen kognitiven Prozess des Bewertens von Informationen im digitalen Raum scheinen gerade junge Menschen Defizite zu haben (Eshet-Alkalai & Amichai-Hamburger, 2004; Eshet-Alkalai & Chajut, 2010; Lorenzen, 2001; van Deursen & van Dijk, 2009). Mit einem multidimensionalen Test, der differenziert für die verschiedenen kognitiven Prozesse ICT-Fähigkeitsschätzungen erfassen und rückmelden kann, könnten solche Defizite im Hinblick auf spezifische kognitive Prozesse identifiziert werden.

Allerdings bleibt mit Blick auf die im Rahmen des Projekts entwickelten Items, trotz dem formulierten Anspruch durch das zugrundeliegende Framework kognitive Aufgaben zu beschreiben, die nicht an spezifische Computerapplikationen gebunden sind und schnell veralten können. Zu konstatieren ist allerdings, dass dies nicht in gleichem Maße für die entwickelten Testitems gilt, die die beschrieben Aufgaben repräsentieren. Die Darstellung der Items ist zum einen bereits stark abstrahiert und in Bezug auf Interaktionsmöglichkeiten eingeschränkt, zum anderen bedürfen sie durch die Veränderungen und Weiterentwicklung von Technologien regelmäßiger inhaltlicher Revision. Beispielsweise ist die in den Items verwendete Desktopoberfläche durch Windows 10 schon zunehmend weniger verbreitet. Auch die Nutzung von anderen Nutzeroberflächen als die von PCs, wie Tablets oder Smartphones, ist durch die Items nicht ausreichend abgebildet, obwohl gerade die Nutzung dieser Endgeräte in besonderem Maße verbreitet ist. Bei der Zielgruppe 15-jähriger Schülerinnen und Schüler wären möglicherweise Befragungen zur Akzeptanz der ICT-Items sinnvoll, um die Aktualität der Items immer wieder kritisch zu überprüfen und gegebenenfalls Aktualisierungen im Layout und/oder der Funktionsweise vornehmen zu können. Darüber hinaus sollten Veränderungen in der Nutzung und Verfügbarkeit von Applikationen durch die rasche Weiterentwicklung digitaler Technologien auch in neuen ICT-Items repräsentiert werden. In den 64 ICT-Items werden momentan vor allem Applikationen wie Browser, E-Mail und gängige Office-Anwendungen abgebildet. In Bezug auf das aktuelle Geschehen ist aber beispielsweise eine gesteigerte Nutzung von Applikationen, wie Video-Konferenz-Tools oder Instant-Messaging, vor allem im schulischen und beruflichen Bereich zu beobachten, die verteiltes Arbeiten und digitale Kommunikation in Gruppen ermöglichen und die durch die vorliegenden ICT-Items nicht adressiert werden. Denkbar ist auch, dass sich in der nahen Zukunft stärker immersive Formen der Nutzung digitaler Technologien weiter durchsetzen. Das CavE-ICT-Framework könnte solche Weiterentwicklungen

theoretisch durch die Facette der Modalität abbilden. Ob diese aber auch künftig in Items repräsentiert sind, ist unklar, da hier Grenzen der Praktikabilität in der Testrealität liegen.

Es wird also deutlich, dass stetige Neu- und Weiterentwicklungen von ICT-Items aus inhaltlicher Sicht erforderlich sind. Hinzu kommt, dass im Rahmen der zweiten Studie dieser Arbeit auch deutlich wurde, dass der bestehende ICT-Itempool für eine Nutzung im Rahmen des multidimensionalen adaptiven Testens zur Rückmeldung differenzierter ICT-Fähigkeiten nach kognitivem Prozess nicht optimal ist. Zum einen würden die erprobten Testalgorithmen generell davon profitieren, wenn insgesamt ein größerer Itempool zur Verfügung stehen würde, aus dem adaptiv ICT-Items ausgewählt werden könnten. Zum anderen wäre eine gezielte Entwicklung von Items erstrebenswert, um eine bessere Verteilung der Items für die kognitiven Prozesse, idealerweise einer Gleichverteilung von Itemzahlen pro kognitivem Prozess, zu erreichen. Zudem wäre es günstig, wenn sich die Items zur Messung der jeweiligen kognitiven Prozesse in ihrer Schwierigkeit auch breiter und gleichförmiger über das ICT-Fähigkeitsspektrum verteilen würden. Die deutlichen Unterschiede zwischen den Hinweisen zur Itempoolentwicklung sowie der antizipierten CAT-Testlänge aus Studie I und den Ergebnissen unter Nutzung von Itemparametern sowie der Verteilung der Personenfähigkeiten des ICT-Feldtests in Studie II, lassen sich vorrangig durch den nicht optimal aufgebauten ICT-Itempool erklären.

Eine stetige Erweiterung und Aktualisierung von Itempools, wie sie für eine erfolgreiche und nachhaltige Nutzung eines ICT-Skills-Tests notwendig erscheint, kann eher im Rahmen von wiederkehrend durchgeführten Studien mit langfristigem Zeithorizont gewährleistet werden. LSAs sind ein Beispiel für solche Studien. Die Nutzung ICT-Skills-Tests die im Rahmen von LSAs entwickelt werden und sich mindestens in Teilen auch aus simulationsbasierten, verhaltensnahen Items zusammensetzen, wie beispielsweise in ICILS oder NEPS (Senkbeil & Ihme, 2019), ist allerdings allein der

Anwendung im jeweiligen LSA-Kontext vorbehalten. Ein Transfer oder eine alternative Zusammenstellung und Administration der entwickelten ICT-Items sowie der Übertrag auf andere Testszenarien im Bereich der Gruppen- und Individualdiagnostik findet nicht statt.

9.2.2 Optimierte Itemauswahl zur Zusammenstellung linearer Tests

In Studie III der vorliegenden Arbeit konnte gezeigt, werden, dass kurze Tests zusammengestellt werden konnten, die aber dennoch das recht komplexe Konstrukt der ICT-Skills gut und umfassend abbildenden konnten. Aus dem vorgegebenen ICT-Itempool wurden Items zu einem Test selektiert, die zum einen die formulierten Testziele erfüllen, dabei aber auch zuvor definierte Nebenbedingungen einhalten. Dabei sind automatisierte Verfahren zur optimalen Itemauswahl eine große Erleichterung. Änderungen in den gesetzten Kriterien zur Testzusammenstellung, beispielsweise Anpassungen der Zielfunktion oder auch eine Erweiterung oder Reduzierung der Nebenbedingungen an die Testzusammenstellung, können leicht und schnell umgesetzt werden. Ist der zur Verfügung stehende Itempool entsprechend groß, könnten so Tests für verschiedene Einsatzbedingungen immer wieder schnell und flexibel abgeleitet werden. Festzuhalten ist allerdings, dass die formalisierte Umsetzung der Zielfunktion und der Nebenbedingungen nicht trivial ist. Für die Nutzung der Vorteile von computerbasiert automatisierten Testzusammenstellung ist auch der Zugang beziehungsweise das Verständnis und die Fähigkeit zu entsprechenden Programmierungen erforderlich. Zudem gibt es neben dem in Studie III genutzten Ansatz nach van der Linden (2005) eine Anzahl anderer Ansätze zur automatisierten Zusammenstellung von Test (siehe Schultze, 2017). Um eine fundierte Entscheidung für die Nutzung eins Ansatzes zur Testzusammenstellung treffen und diesen Ansatz sinnvoll umsetzen zu können ist es in jedem Fall nötig sich zunächst methodisch einzuarbeiten. Damit können automatisierte

Testzusammenstellung zwar schnell angepasst werden, allerdings ist der vorgelagerte zeitliche Aufwand zur Einarbeitung in die Methode nicht zu unterschätzen.

9.2.3 Zugrundeliegendes Messmodell

Das im Rahmen des CavE-ICT-Projekts zugrunde gelegte eindimensionale IRT-Modell wies im Vergleich zu einem MIRT-Modell (Adams, Wilson & Wang, 1997) eine bessere Passung zu den vorliegen Daten auf. Itemparameter, die unter Nutzung des eindimensionalen IRT-Modelles geschätzt wurden, wurden für eine differenzierte Rückmeldung von ICT-Skills für die fünf kognitiven Prozesse fixiert und unter Nutzung eines fünf-dimensionalen Rasch-Modells wurden Personenfähigkeiten geschätzt. Auch wenn ein solches Vorgehen im Rahmen von LSAs durchaus gängig ist, gibt es an diesem Ansatz auch Kritik, sofern ein eindimensionales Konstrukt postuliert, dann aber doch mehrere Dimensionen angenommen und als Berichtsmetrik verwendet werden (Brandt, 2015). Wenn das Ziel ist, Rückmeldungen zu Fähigkeiten für ein Merkmal und gleichzeitig auch für einzelne Merkmalsdimensionen oder Subfacetten des Merkmals zu ermöglichen, könnten beispielsweise auch Higher-Order und hierarchische IRT-Modelle oder ein Subdimensionsmodell eingesetzt werden, die dies in besonderer Weise berücksichtigen (Brandt, 2015).

Betrachtet man die entwickelten ICT-Items kritisch, fällt auf, dass die Zuordnung der Items zu einem der fünf kognitiven Prozesse nicht immer unumstritten und eindeutig ausfällt. Obwohl immer ein kognitiver Prozess für die Lösung der spezifischen ICT-Aufgabe vordergründig ist, werden meist weitere Prozesse benötigt, um zur richtigen Lösung zu gelangen. Dies würde dafür sprechen ein MIRT-Modell mit Within-Item-Multidimensionalität zu nutzen, bei dem die Lösung eines Items mehrerer Merkmalsdimensionen bedarf (bspw. Reckase, 2009). Im Rahmen des CavE-ICT-Projekts und dieser Arbeit wurden allerdings nur MIRT-Modelle mit Between-Item-Multidimensionalität verwendet, die die Lösung eines Items durch genau eine

Merkmalsdimension vorhersagen. Bisher existieren nur wenige Anwendungen von Within-Item-Modellen (bspw. Hartig & Höhler, 2010). Die Exploration der Möglichkeit zur Nutzung von derartigen MIRT-Modellen für die Messung von ICT-Skills wäre angezeigt. Zudem könnten diese auch im Rahmen des MAT für Itemauswahl und Fähigkeitsschätzung verwendet werden.

9.2.4 Multidimensionales adaptives Testen

In Studie I und II der vorliegenden Arbeit wurden adaptive Testalgorithmen verglichen, die Multidimensionalität in verschiedener Weise berücksichtigten. Generell kann Multidimensionalität auf zwei Arten in den Testalgorithmus aufgenommen werden: 1) bei der vorläufigen Fähigkeitsschätzung und der auf dieser Basis vorgenommenen adaptiven Itemauswahl; 2) bei der abschließenden Fähigkeitsschätzung (Kröhne et al., 2014). Anhand der in dieser Arbeit betrachteten adaptiven Algorithmen konnte gezeigt werden, dass die Art der Itemauswahl (nach maximaler Information) und der vorläufigen Fähigkeitsschätzung im CAT-Verlauf keinen so großen Effekt auf die Messgenauigkeit hat wie die abschließende multidimensionale Fähigkeitsschätzung. Ist der adaptive Test hinreichend lang, sind kaum bedeutsame Unterschiede zwischen den verschiedenen Algorithmen zu erkennen. Sollen daher – wie im Falle der implementierten C-MATs – unerwünschte Effekte durch eine vermischte Vorgabe von Items verschiedener Dimensionen vermieden werden, kann bei der Testbearbeitung auf einen in der Itemauswahl stärker einschränkenden Algorithmus zurückgegriffen werden. Noch stärker als bei dem im Rahmen dieser Arbeit im Fokus stehenden Konstrukt der ICT-Skills kann dies bei einer simultanen Messung mehrerer korrelierte Merkmale durch MAT von Vorteil sein (Kröhne et al., 2014).

Generell gilt, dass restriktivere Itemauswahlmechanismen angewendet werden könnten, sofern der Itempool ausreichend groß ist und die abschließende Schätzung der Personenfähigkeiten die Multidimensionalität einbezieht.

9.2.5 Erkenntnisgewinn durch Simulationsstudien

Im Rahmen dieser Arbeit wurden drei Simulationsstudien durchgeführt. Die Nutzung von Simulationsstudien in unterschiedlichen Phasen der Testentwicklung ist durchaus angezeigt (Thompson & Weis, 2011). Sie können helfen, noch vor Beginn der Test- bzw. Itementwicklung zu prüfen, ob für das postulierte Konstrukt eine differenzierte Messung möglich ist, wieviele Items im zu entwickelnden Itempool vorliegen sollen, wie dieser Itempool strukturiert sein sollte oder wie lang ein späterer Test voraussichtlich werden würde, um eine reliable und valide Messung vornehmen zu können. Dabei sollte beachtet werden, dass die Rahmenbedingungen in Simulationsstudien so gesetzt werden sollten, dass sie möglichst nah an der tatsächlichen Testrealität sind. Wenn nur wenig über die Rahmenbedingungen bekannt ist, kann es passieren, dass aufgrund der großen Abstraktion und ungenügenden Einplanung von Ungleichheiten oder Abweichungen nicht genügend belastbare Informationen für die eigentliche Testentwicklung abgeleitet werden können. Zwar können die Ergebnisse früher Simulationsstudien wichtige Hinweise liefern, sie sollten allerdings eher als optimistische Einschätzungen verstanden und nur als grobe Richtlinie genutzt werden.

Viele Simulationsstudien im Bereich des adaptiven Testens beziehen sich auf große Itempools und/oder große Stichproben. Sie werden meist mit bereits bestehenden Itempools durchgeführt, deren Weiterentwicklung von linearen zu adaptiven Instrumenten forciert wird oder versuchen die Funktionsweise neuer methodischer Entwicklungen darzustellen (bspw. Born & Frey, 2017; ein Überblick zu Simulationsstudien die lineare Tests, CAT- und MAT-Algorithmen vergleichen findet sich bspw. bei Frey & Seitz, 2009).

Neue Testentwicklungen, die direkt eine adaptive Vorlage in den Blick nehmen, noch bevor ein Item entwickelt wurde, sind dagegen eher selten. Meist werden zudem vorrangig einfachere Itemformate umgesetzt, sodass mit weniger Aufwand größere Itempools

erstellt werden können, als dies für das in dieser Arbeit im Fokus stehende Konstrukt der ICT-Skills der Fall ist. Komplexe Konstrukte, die noch dazu über komplexe simulationsbasierte Items erfasst werden, bilden hier eine spezielle Herausforderung. Kleine Itempools und verhältnismäßig kleine Stichproben werden in den meisten Simulationsstudien nicht antizipiert. Daneben sollten auch unausgewogene Verteilungen von Items über das Schwierigkeitsspektrum oder auf mehrere Merkmalsdimensionen beachtet werden. Außerdem könnten unterschiedlich hohe Korrelationen zwischen diesen einzelnen Dimensionen in entsprechenden Simulationsstudien stärker berücksichtigt werden. Insgesamt sollte eine größere Variabilität in die generierten Daten aufgenommen werden.

Simulationen, die Echtdaten einbeziehen (Hybridsimulationen), sollten im Prozess der Testentwicklung möglichst früh durchgeführt werden, um Weiterentwicklungen gerichtet vorantreiben und genauere Abschätzungen vornehmen zu können. Die in Studie II vorgenommen Simulation bildet die Realität sehr viel klarer ab, zeigt aber auch ein weniger optimistisches Bild für die Testanwendung als dies durch Studie I nahegelegt wurde. In den Studien I und II zeigen sich deutliche Unterschiede hinsichtlich des auf Basis der Ergebnisse empfohlenen Tests, vor allem bezüglich der Testlänge, aber auch im Hinblick auf die adaptiven Algorithmen.

9.3 Fazit

ICT-Skills stellen ein hoch komplexes Konstrukt dar, das im Rahmen des CavE-ICT-Frameworks durch eine Vielzahl von Facetten und Facettenebenen beschrieben werden kann. Die Testitems, die ICT-spezifische Aufgaben simulationsbasiert und verhaltensnah erfassen sollen, sind sehr aufwändig in ihrer Entwicklung, weswegen letztlich im ICT-Itempool auch eine verhältnismäßig geringe Anzahl an Items vorliegt. Zudem sind die schließlich entwickelten ICT-Items hinsichtlich ihrer mittleren

Bearbeitungszeit sehr heterogen und veranschlagen auch insgesamt betrachtet relativ hohe Bearbeitungszeiten, was in der Folge auch die Testzeiten in die Höhe treibt und somit eine ökonomische und zugleich zuverlässige Messung von ICT erschwert. In Anbetracht dessen können folgende Kernerkenntnisse aus der vorliegenden Arbeit abgeleitet werden:

1. Wenn eine differenzierte und reliable Rückmeldung zu Facettenebenen des Konstrukts gegeben werden soll, steht man vor einer großen Herausforderung. MAT löst das Problem nur zum Teil da die Testzeiten weiterhin recht hoch ausfallen. Was allerdings gezeigt werden konnte, ist, dass eine sequentielle Vorgabe von ICT-Items nach kognitiven Prozessen möglich ist, ohne viel Messpräzision oder Messeffizienz einzubüßen. Diese Erkenntnis lässt sich auch auf die Erfassung anderer multidimensionaler Konstrukte oder korrelierter Merkmale übertragen, die gemeinsam in einem MAT erfasst werden sollen. Die Form der Itemauswahl im CAT-Verlauf scheint zudem keinen allzu großen Effekt auf die Messpräzision zu haben. Weit relevanter für eine präzise und zuverlässige Messung sind die zur abschließenden Fähigkeitsschätzung genutzten Zusammenhänge der Merkmalsdimensionen.
2. Simulationsstudien sind im Prozess der Testentwicklung ein unverzichtbares Werkzeug, um bereits früh im Prozess Annahmen und Möglichkeiten zum späteren Testeinsatz zu prüfen und gegebenenfalls Anpassungen im Entwicklungsprozess oder bezüglich der Zielsetzung des Tests vorzunehmen. Dennoch bleibt festzuhalten, dass die Rahmenbedingungen in Simulationen möglichst realistisch, vielleicht sogar eher pessimistisch gesetzt werden sollten, um wirklich belastbare oder wenigstens konservativere Prognosen daraus ableiten zu können.
3. Vor allem bei so aufwändigen, wie in dieser Arbeit vorgestellten, Entwicklungen von Items ist es lohnenswert, Möglichkeiten unterschiedlicher Testszenarien und

Anwendungsfelder zu prüfen. Optimierte und automatisierte Itemauswahl, um beispielsweise Kurztests aus einem größeren Itempool abzuleiten, bieten die Möglichkeit, recht einfach und schnell Anpassungen in der Zielsetzung oder in den Nebenbedingungen vorzunehmen. So kann der unter großem Aufwand erstellte ICT-Itempool auch für andere – individual- und gruppendiagnostische – als für die zunächst intendierten, populationsbeschreibenden Zwecke genutzt werden.

Literaturverzeichnis

Adams, R. J. & Wilson, M. (1996). Formulating the Rasch Model as a mixed coefficients multinomial logit. In G. Engelhard, JR. & M. Wilson (Hrsg.), *Objective measurement: Theory into practice* (Bd. 3, S. 143–166). Norwood, NJ: Ablex Publishing Company.

Adams, R. J., Wilson, M. & Wang, W.-C. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, *21*, 1–23. https://doi.org/10.1177/0146621697211001

Akaike, H. (1973). Information Theory as an extension of the Maximum Likelihood Principle. In B. N. Petrov & F. Csaki (Hrsg.), *Second international symposium on Information Theory* (S. 267–281). Budapest: Akaemiai Kiado.

Alheit, P. & Dausien, B. (2002). Bildungsprozesse über die Lebensspanne und lebenslanges Lernen. In R. Tippelt (Hrsg.), *Handbuch Bildungsforschung* (S. 565–585). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-99634-3_31

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrew Makhorin. (2012). GNU Linear Programming Kit [Computer software]. Verfügbar unter https://www.gnu.org/software/glpk/

Babcock, B. & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, *36*, 565–580. https://doi.org/10.1177/0146621612455090

Babcock, B. & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests. Variable-length CATs are not biased. In D. J. Weiss (Hrsg.), *Proceedings of the 2009 GMAC Conference on computerized adaptive testing*. Verfügbar unter www.psych.umn.edu/psylabs/CATCentral/

Baumert, J., Stanat, P. & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 15–68). Opladen: Leske + Budrich.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M. et al. (2012). Defining twenty-first century skills. In P. Griffin, McGaw Barry & E. Care (Hrsg.), *Assessment and teaching of 21st century skills* (S. 17–66). Dordrecht: Springer.

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

Blossfeld, H.-P. & Roßbach, H.-G. (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS). Edition ZfE.* (2. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bloxom, B. & Vale, C. D. (1987). *Multidimensional adaptive testing: An approximate procedure for updating.* Paper presented at the Meeting of the Psychometric Society, Montreal.

Bock, R. D. & Aitkin, M. (1981). Marginal Maximum Likelihood Estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. Verfügbar unter https://link.springer.com/article/10.1007/BF02293801

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/10.1177/014662168200600405

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates, 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

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler* (4. Aufl.). Heidelberg: Springer.

Bos, W., Eickelmann, B., Gerick, J., Goldhammer, F., Schaumburg, H., Schwippert, K. et al. (Hrsg.). (2014). *ICILS 2013. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern in der 8. Jahrgangsstufe im internationalen Vergleich*. Münster: Waxmann.

Brandt, S. (2015). *Unidimensional interpretation of multidimensional tests*. Dissertation. Christian-Albrechts-Universität zu Kiel. Retrieved from https://macau.uni-kiel.de/

Brandt, S. & Duckor, B. (2013). Increasing unidimensional measurement precision using a multidimensional item response model approach. *Psychological Test and Assessment Modeling, 55*, 148–161. Retrieved from https://www.psychologie-aktuell.com

Bundesministerium für Bildung und Forschung. (2016). *Bildungsoffensive für die digitale Wissensgesellschaft. Strategie des Bundesministeriums für Bildung und Forschung* (Referat Digitaler Wandel in der Bildung, Hrsg.). Berlin.

Calvani, A., Cartelli, A., Fini, A. & Ranieri, M. (2008). Models and Instruments for assessing Digital Competence at School. *Journal of E-Learning and Knowledge Society*, *4*, 183–193. https://doi.org/10.20368/1971-8829/288

Carretero, S., Vuorikari, R. & Punie, Y. (2017). *DigComp 2.1: The digital competence framework for citizens with eight proficiency levels and examples of use*. Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/38842

Chalmers, R. P. (2012). MIRT: A multidimensional item response theory package for the R environment (Version 1.23) [Computer software]. Verfügbar unter https://www.jstatsoft.org/article/view/v048i06

Chandler, P. & Sweller, J. (1991). Cognitive Load Theory and the format of instruction. *Cognition and Instruction*, *8*, 293–332. Verfügbar unter https://www.jstor.org/stable/3233596

Chang, H.-H. & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213–229. https://doi.org/10.1177/014662169602000303

Chen, S.-Y., Ankenmann, R. D. & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*, 129–145. https://doi.org/10.1111/j.1745-3984.2003.tb01100.x

Cheng, P. E. & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement*, *24*, 257–265. https://doi.org/10.1177/01466210022031723

Cheng, Y. & Chang, H.-H. (2009). The Maximum Priority Index Method for severely constrained item selection in computerized adaptive testing. *The British Journal of Mathematical and Statistical Psychology*, *62*, 369–383. https://doi.org/10.1348/000711008X304376

Choi, S. W., Moellering, K. T., Li, J. & van der Linden, W. J. (2016). Optimal reassembly of shadow tests in CAT. *Applied Psychological Measurement*, *40*, 469–485. https://doi.org/10.1177/0146621616654597

Cox, A. M., Vasconcelos, A. C. & Holdridge, P. (2010). Diversifying assessment through multimedia creation in a nontechnical module: Reflections on the MAIK project. *Assessment & Evaluation in Higher Education*, *35*, 831–846. https://doi.org/10.1080/02602930903125249

Danner, D. (2015). *Reliabilität - die Genauigkeit einer Messung*. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines). Verfügbar unter http://dx.doi.org/10.15465/GESIS-SG_011 https://doi.org/10.15465/GESIS-SG_011

Davis, J. & Ferdous, A. (2005). *Using item difficulty and item position to measure test fatigue* (American Institutes for Research, Hrsg.). Washington, D.C. Verfügbar unter www.air.org

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

Desjardins, C. D. & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R* (The R Series). Boca Raton, FL: CRC Press.

Diao, Q. & Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In D. J. Weiss (Hrsg.), *Proceedings of the 2009 GMAC Conference on computerized adaptive testing* .

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–25). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Eccles, J. S. (2006). Families, schools, and developing achievement-related motivations and engagement. In J. E. Grusec & P. D. Hastings (Hrsg.), *Handbook of socialization. Theory and research* (S. 665–691). New York: Guilford Press.

Edmunds, A. & Morris, A. (2000). The problem of information overload in business organisations: A review of the literature. *International Journal of Information Management*, *20*, 17–28. https://doi.org/10.1016/S0268-4012(99)00051-1

Eggen, T. J. H. M. (2008). Adaptive testing and item banking. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 215–234). Göttingen: Hogrefe & Huber Publishers.

Ehmke, T. & Siegle, T. (2008). Einfluss elterlicher Mathematikkompetenz und familialer Prozesse auf den Kompetenzerwerb von Kindern in Mathematik. *Psychologie in Erziehung und Unterricht*, *55*, 253–264. Verfügbar unter https://www.reinhardt-journals.de/index.php/peu/article/view/640

Eickelmann, B. & Bos, W. (2011). Messung computer- und informationsbezogener Kompetenzen von Schülerinnen und Schülern als Schlüsselkompetenz im 21. Jahrhundert. *Medienimpulse*, *49*(4), 1–16.

Eickelmann, B., Bos, W., Gerick, J., Goldhammer, F., Schaumburg, H., Schwippert, K. et al. (Hrsg.). (2019). *ICILS 2018 #Deutschland. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking*. Münster: Waxmann.

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion* (1. Aufl.). Göttingen: Hogrefe Verlag.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *The American Psychologist*, *61*, 50-55. https://doi.org/10.1037/0003-066X.61.1.50

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah: Lawrence Erlbaum Associates, Publishers.

Engelhardt, L., Goldhammer, F., Naumann, J. & Frey, A. (2017). Experimental validation strategies for heterogeneous computer-based assessment items. *Computers in Human Behavior*, *76*, 683–692. https://doi.org/10.1016/j.chb.2017.02.020

Engelhardt, L., Naumann, J., Goldhammer, F., Frey, A., Horz, H., Hartig, K. et al. (eingereicht). Development and evaluation of a framework for the performance-based testing of ICT skills. *Frontiers in Psychology*.

Engelhardt, L., Naumann, J., Goldhammer, F., Frey, A., Wenzel, S. F. C., Hartig, K. et al. (2019). Convergent evidence for the validity of a performance-based ICT skills test. *European Journal of Psychological Assessment*, 1–11. https://doi.org/10.1027/1015-5759/a000507

Eshet Alkali, Y. & Amichai-Hamburger, Y. (2004). Experiments in digital literacy. *CyberPsychology & Behaviour*, *7*, 421–429. https://doi.org/10.1089/cpb.2004.7.421

Eshet-Alkalai, Y. (2004). Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of Educational Multimedia and Hypermedia*, *13*, 93–106. Verfügbar unter https://www.learntechlib.org/primary/p/4793/

Eshet-Alkalai, Y. & Chajut, E. (2010). You can teach old dogs new tricks: The factors that affect changes over time in digital literacy. *Journal of Information Technology Education*, *9*, 173–181.

Ezziane, Z. (2007). Information technology literacy: Implications on teaching and learning. *Journal of Educational Technology & Society*, *10*(3), 175–191. Verfügbar unter https://www.jstor.org/stable/10.2307/jeductechsoci.10.3.175

Ferrari, A., Punie, Y. & Redecker, C. (2012). Understanding digital competence in the 21st century: An analysis of current frameworks. In A. Ravenscroft, S. Lindstaedt, C. Delgado Kloos & D. Hernández-Leo (Hrsg.), *21st century learning for 21st century skills. 7th European Conference on Technology Enhanced Learning, EC-TEL 2012* (S. 79–92). Heidelberg: Springer.

Fraillon, J. & Ainley, J. (2010). *The IEA International Study of computer and information literacy (ICILS)*. Verfügbar unter https://www.researchgate.net/publication/268297993_The_IEA_International_Study_of_Computer_and_Information_Literacy_ICILS

Fraillon, J., Schulz, W. & Ainley, J. (2013). *International computer and information literacy study. Assessment framework* (International Association for the Evaluation of Educational Achievement (IEA), Hrsg.). Amsterdam.

Frey, A. (in Druck). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebodenkonstruktion*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 275–292). Berlin: Springer.

Frey, A. & Annageldyev, M. (2015). Youden. A program for the construction of booklet designs (Version 1.0) [Computer software].

Frey, A., Cheng, Y. & Seitz, N.-N. (2011). *Content balancing with the Maximum Priority Index Method in multidimensional adaptive testing*. Paper presented at the 2011 Meeting of the National Council on Measurement in Education (NCME), New Orleans, LA, USA.

Frey, A. & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), Kompetenzdiagnostik. *Zeitschrift für Eriehungswissenschaft. 10* (8), 169–184 [Themenheft]. Verfügbar unter https://link.springer.com

Frey, A. & Hartig, J. (2013). Wann sollten computerbasierte Verfahren zur Messung von Kompetenzen anstelle von papier- und bleistift-basierten Verfahren eingesetzt werden? *Zeitschrift für Erziehungswissenschaft*, *16*, 53–57. https://doi.org/10.1007/s11618-013-0385-1

Frey, A., Hartig, J. & Rupp, A. A. (2009). An NCME instructional module on Booklet Designs in Large-Scale Assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53. https://doi.org/10.1111/j.1745-3992.2009.00154.x

Frey, A. & Seitz, N.-N. (2009). Multidimensional adaptive testing in educational and psychological measurement. Current state and future challenges. *Studies in Educational Evaluation*, *35*, 89–94. https://doi.org/10.1016/j.stueduc.2009.10.007

Frey, A. & Seitz, N.-N. (2010). Multidimensionale adative Kompetenzdiagnostik: Ergebnisse zur Messeffizienz. Projekt MAT. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), Kompetenzmodellierung. Zwischenbilanz des DFG Schwerpunktprogramms und Perspektiven des Forschungsansatzes. *Zeitschrift für Pädagogik*, 40–51 [Themenheft]. Weinheim: Beltz.

Frey, A., Seitz, N.-N. & Kröhne, U. (2013). Reporting differentiated literacy results in PISA by using multidimensional adaptive testing. In M. Prenzel, M. Kobarg, K. Schöps & S. Rönnebeck (Hrsg.), *Research on PISA* (S. 103–120). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-4458-5_7

Georgiadou, E., Triantafillou, E. & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment*, *5*(8), 1–39. Verfügbar unter http://www.jtla.org.

Glas, C. A. W. (2010). Item Parameter Estimation and Item Fit Analysis. In W. J. van der Linden & C. A. W. Glas (Hrsg.), *Elements of adaptive testing* (S. 269–288). New York: Springer. https://doi.org/10.1007/978-0-387-85461-8_14

Gniewosz, B. & Gräsel, C. (2015). VIII-1 Überblick Soziale Ungleichheit. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Hrsg.), *Empirische Bildungsforschung. Gegenstandsbereiche* (2. Aufl., S. 195–200). Wiesbaden: VS Verlag für Sozialwissenschaften.

Goldhammer, F., Gniewosz, G. & Zylka, J. (2016). ICT engagement in learning environments. In S. Kuger, E. Klieme, N. Jude & D. Kaplan (Hrsg.), *Assessing contexts of learning; An international perspective.* (S. 331–351). Cham: Springer.

Goldhammer, F. & Hartig, J. (2012). Interpretation von Testresultaten und Testeichung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch, 2., aktualisierte und überarbeitete Auflage, S. 173–201). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20072-4_8

Goldhammer, F., Kröhne, U., Keßel, Y., Senkbeil, M. & Ihme, J. M. (2014). Diagnostik von ICT-Literacy. *Diagnostica*, *60*, 10–21. https://doi.org/10.1026/0012-1924/a000113

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F. & Funke, J. (2013). Computer-based assessment of complex problem solving: Concept, implementation, and application. *Educational Technology Research and Development*, *61*, 407–421. https://doi.org/10.1007/s11423-013-9301-x

Gu, L. & Reckase, M. D. (2007). Designing optimal item pools for computerized adaptive tests with Sympson-Hetter exposure control. In D. J. Weiss (Hrsg.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. .

Hambleton, R. K., Zaal, J. N. & Pieters, J. P. M. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K. Hambleton & J. N. Zaal (Hrsg.), *Advances in educational and psychological testing: Theory and applications* (S. 341–366). New York: Kluwer Academic/Plenum.

Hambleton, R. K. & Swaminathan, H. (2010). *Item response theory. Princeples and applications* (Evaluation in education and human services series). Boston: Kluwer. https://doi.org/10.1007/978-94-017-1988-9

Hartig, J. & Goldhammer, F. (2010). Modelle der Item-Response-Theorie. In S. Maschke & L. Stecher (Hrsg.), *Enzyklopädie Erziehungswissenschaft Online. Fachgebiet: Methoden der empirischen erziehungswissenschaftlichen Forschung, Quantitative Forschungsmethoden* (S. 1–36). Weinheim: Juventa Verlag. https://doi.org/10.3262/EEO07100111

Hartig, J. & Höhler, J. (2010). *Modellierung von Kompetenzen mit mehrdimensionalen IRT-Modellen. Projekt MIRT*. Weinheim: Beltz.

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik. Mit 18 Tabellen* (S. 127–143). Berlin, Heidelberg: Springer Medizin Verlag Heidelberg. https://doi.org/10.1007/3-540-33020-8_9

He, W., Diao, Q. & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. *Educational and Psychological Measurement*, *74*, 677–696. https://doi.org/10.1177/0013164413517503

Heckhausen, J. & Heckhausen, H. (2010). Motivation und Entwicklung. In J. Heckhausen & H. Heckhausen (Hrsg.), *Motivation und Handeln* (S. 427–486). Berlin: Springer.

Horz, H. (2009). Medien. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (Springer-Lehrbuch, S. 103–133). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-540-88573-3_5

Horz, H. & Schnotz, W. (2010). Cognitive load in learning with multiple representations. In J. L. Plass, R. Moreno & R. Brünken (Hrsg.), *Cognitive Load Theory* (S. 229–252). Campridge: Campridge University Press.

Horz, H., Winter, C. & Fries, S. (2009). Differential benefits of situated instructional prompts. *Computers in Human Behavior*, *25*, 818–828. https://doi.org/10.1016/j.chb.2008.07.001

International ICT Literacy Panel. (2002). *Digitial transformation: A framework for ICT literacy* (Educational Testing Service, Hrsg.). Princeton, NJ.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359–375. https://doi.org/10.1207/s15324818ame0204_6

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. Verfügbar unter https://www.pedocs.de/volltexte/2011/4493/pdf/ZfPaed_2006_Klieme_Leutner_Kompetenzmodelle_Erfassung_Lernergebnisse_D_A.pdf

Kröhne, U. & Frey, A. (2014). *Multidimensional Adaptive Testing Environment (MATE) - Manual.*

Kröhne, U., Goldhammer, F. & Partchev, I. (2014). Constrained multidimensional adaptive testing without intermixing items from different dimensions. *Psychological Test*

and Assessment Modeling, *56*, 348–367. Verfügbar unter https://www.psychologie-aktuell.com

Kubinger, K. (2014). Large-Scale Assessment. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (18. Aufl., S. 915). Bern: Hogrefe.

Kuger, S., Linberg, T., Bäumer, T. & Struck, O. (2018). *Digitale Lernumwelten (NEPS Survey Paper No. 32)* (Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel, Hrsg.). Bamberg.

Kultusminister Konferenz. (2016). Bildung in der digitalen Welt. Strategien der Kultusministerkonferenz. Verfügbar unter https://www.kmk.org/dokumentation-statistik/beschluesse-und-veroeffentlichungen/bildung-in-der-digitalen-welt.html

La Torre, J. de & Patz, R. J. (2005). Making the most of what we have. A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*, 295–311. https://doi.org/10.3102/10769986030003295

Leibniz-Institut für Bildungsverläufe e.V. (2018). *Informationen zur Kompetenztestung NEPS Startkohorte 4 - Klasse 9 Schule und Ausbildung - Bildung von Schülerinnen und Schülern ab Klassenstufe 9*. Bamberg.

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/10.1177/0013164413486802

Leutner, D., Klieme, E., Fleischer, J. & Kuper, H. (2013). Editorial: Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. *Zeitschrift für Erziehungswissenschaft*, *16*, 1–4. https://doi.org/10.1007/s11618-013-0378-0

Li, Y. H. & Schafer, W. D. (2005). Increasing the homogeneity of CAT's item-exposure rates by minimizing or maximizing varied target functions while assembling shadow

tests. *Journal of Educational Measurement*, *42*, 245–270. https://doi.org/10.1111/j.1745-3984.2005.00013.x

Linacre, J. M. (1994). *Many-Facet Rasch Measurement*. Chicago: MESA Press.

Linacre, J. M. (2000). Computer-adaptive testing: A methodolgy whose time has come. In C. Sunhee, K. Unson, J. Eunhwa & J. M. Linacre (Hrsg.), *Development of computerized middle school achievement test*. Seoul, South Korea: Komesa.

Livingstone, S. & Helsper, E. (2007). Gradations in digital inclusion: Children, young people and the digital divide. *New Media & Society*, *9*, 671–696. https://doi.org/10.1177/1461444807080335

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

Lord, F. M. (1986). Maximum Likelihood and Bayesian Parameter Estimation in Item Response Theory. *Journal of Educational Measurement*, *23*, 157–162. Verfügbar unter https://www.jstor.org/stable/1434513

Lorenzen, M. (2001). The land of confusion? *Research Strategies*, *18*(2), 151–163. https://doi.org/10.1016/S0734-3310(02)00074-5

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. *Psychologische Rundschau*, *58*, 103–117. https://doi.org/10.1026/0033-3042.58.2.103

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*, 389–404. Verfügbar unter https://conservancy.umn.edu

Markauskaite, L. (2007). Exploring the structure of trainee teachers' ICT literacy: the main components of, and relationships between, general cognitive and technical capabilities. *Educational Technology Research and Development*, *55*(6), 547–572. https://doi.org/10.1007/s11423-007-9043-8

Masters, G. N. (1982). A Rasch Model for partial credit scoring. *Psychometrika*, *47*, 149–174. https://doi.org/10.1007/BF02296272

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, *58*, 2078–2091. https://doi.org/10.1002/asi.20672

Mikolajetz, A. (2017). *Messung komplexer Kompetenzkonstrukte in Large-Scale-Assessments mit Hilfe von multidimensionalen adaptiven Testens*. Dissertation. Friedrich-Schiller-Universität, Jena.

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine*, *178*(10 Suppl), 107–114. https://doi.org/10.7205/MILMED-D-13-00213

Moosbrugger, H. (2012). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 227–252). Berlin: Springer.

Naumann, J., Richter, T. & Groeben, N. (2001). Validierung des INCOBI anhand eines Vergleichs von Anwendungsexperten und Anwendungsnovizen. *Zeitschrift für Pädagogische Psychologie*, *15*, 219–232. https://doi.org/10.1024//1010-0652.15.34.219

Nydick, S. W. & Weiss, D. J. (2009). A hybrid simulation procedure for the development of CATs. In D. J. Weiss (Hrsg.), *Proceedings of the 2009 GMAC Conference on computerized adaptive testing* .

OECD. (2009). *PISA 2006 Technical Report* (PISA). Paris: OECD Publishing.

OECD. (2012). *PISA 2009 Technical Report* (PISA). Paris: OECD Publishing.

OECD. (2014). *PISA 2012 Technical Report* (PISA). Paris: OECD Publishing.

OECD. (2018). *PISA 2015 Technical Report* (PISA). Paris: OECD Publishing.

Osterlind, S. J. & Everson, H. T. (2009). *Differential item functioning* (2. Aufl.). Los Angeles, CA: Sage.

Parshall, C. G., Spray, J. A., Kalohn, J. & Davey, T. *Practical considerations in computer-based testing*. New York: Springer.

Poynton, T. A. (2005). Computer literacy across the lifespan: A review with implications for educators. *Computers in Human Behavior*, *21*, 861–872. https://doi.org/10.1016/j.chb.2004.03.004

R Core Team. (2016). R (Version 3.3.1) [Computer software]. Verfügbar unter https://www.r-project.org/

Rammstedt, B. (Hrsg.). (2013). *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich. Ergebnisse von PIAAC 2012*. Münster: Waxmann. Verfügbar unter https://nbn-resolving.org/urn:nbn:de:0168-ssoar-360687

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen: The Danish Institute for Educational Research.

Reckase, M. D. (2009). *Multidimensional item response theory* (Statistics for Social and Behavioral Sciences). Dordrecht: Springer. https://doi.org/10.1007/978-0-387-89976-3

Richter, T., Naumann, J. & Horz, H. (2010). Eine revidierte Fassung des Inventars zur Computerbildung (INCOBI-R). *Zeitschrift für Pädagogische Psychologie*, *24*, 23–37. https://doi.org/10.1024/1010-0652/a000002

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*, *53*, 145–161. https://doi.org/10.1002/asi.10017

Robitzsch, A., Kiefer, T. & Wu, M. (2017). TAM: Test analysis modules (Version 2.8-21) [Computer software]. Verfügbar unter https://CRAN.R-project.org/package=TAM

Rölke, H. (2012). The ItemBuilder: A graphical authoring system for complex item development. In T. Bastiaens & G. Marks (Hrsg.), *Proceedings of World Conference on E-Learning in Corporate, Gouvernmet, Healthcare, and Higher Education* (S. 344–353). Montréal: Association for the Advancement of Computing in Education (AACE). Verfügbar unter https://www.learntechlib.org/p/41614/

Rost, J. (2004). *Lehrbuch Testtheorie -Testkonstruktion* (2. Aufl.). Bern: Huber.

Rost, J., Prenzel, M., Carstensen, C. H., Senkbeil, M. & Groß, K. (2004). *Naturwissenschaftliche Bildung in Deutschland. Methoden und Ergebnisse von PISA 2000*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Samejima, F. (1996). *Estimation of latent ability using a response pattern of graded scores*. Richmond, VA: The William Byrd Press.

Schermelleh-Engel, K. & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 119–140). Berlin: Springer.

Schmeiser, C. B. & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (ACE /Praeger Series on Higher Education, 4th ed., pp. 307–353). Westport, CT: Praeger Publ.

Schmidt-Atzert, L., Amelang, M. & Fydrich, T. (2012). *Psychologische Diagnostik. Mit 82 Tabellen* (Springer-Lehrbuch, 5., vollständig überarbeitete und erweiterte Auflage). Berlin: Springer. https://doi.org/10.1007/978-3-642-17001-0

Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Hrsg.), *The Cambridge handbook of multimedia learning* (S. 49–70). New York: Cambridge University Press.

Schultze, M. (2017). *Constructing subtests using Ant Colony Optimization*. Dissertation. Freie Universität Basel.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. Verfügbar unter https://www.jstor.org/stable/2958889

Schweizer, K., Schreiner, M. & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two-dimensional model of APM. *Psychology Science Quarterly*, *51*, 47–64. Verfügbar unter https://www.researchgate.net

Segall, D. O. (2005). Computerized Adaptive Testing. In K. Kempf-Leonard (Hrsg.), *Encyclopedia of social measurement* (S. 429–438). Amsterdam: Elsevier.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354. Verfügbar unter https://link.springer.com

Senkbeil, M. (2018). Development and validation of the ICT motivation scale for young adolescents. Results of the international school assessment study ICILS 2013 in Germany. *Learning and Individual Differences*, *67*, 167–176. https://doi.org/10.1016/j.lindif.2018.08.007

Senkbeil, M. & Ihme, J. M. (2017). Entwicklung und Validierung eines Kurzfragebogens zur Erfassung computerbezogener Anreizfaktoren bei Erwachsenen. *Diagnostica*, *63*(2), 87–98. https://doi.org/10.1026/0012-1924/a000170

Senkbeil, M. & Ihme, J. M. (2019). Diagnostik von ICT Literacy: Messen Multiple-Choice-Aufgaben und simulationsbasierte Aufgaben vergleichbare Konstrukte? *Diagnostica*, 1–11. https://doi.org/10.1026/0012-1924/a000243

Shin, C. D., Chien, Y., Way, W. D. & Swanson, L. (2009). *Weighted penalty model for content balancing in CATs*. San Antonio, TX: Pearson.

Sireci, S. G. & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Stocking, M. L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*, 277–292. https://doi.org/10.1177/014662169301700308

Sympson, J. B. & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In Military Testing Association (Hrsg.), *Proceedings of the 27th annual*

meeting of the Military Testing Association (S. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

Thompson, N. A. & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16*(1), 1–9. Verfügbar unter https://pareonline.net/getvn.asp?v=16&n=1

Van der Linden, W. J. (1998a). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*, 201–216. Verfügbar unter https://www.researchgate.net

Van der Linden, W. J. (1998b). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*, 195–211. https://doi.org/10.1177/01466216980223001

Van der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, *23*, 21–29. https://doi.org/10.1177/01466219922031149

Van der Linden, W. J. (2005). *Linear Models for Optimal Test Design* (Statistics for Social and Behavioral Sciences). New York, NY: Springer Science+Business Media Inc. https://doi.org/10.1007/0-387-29054-0

Van der Linden, W. J. (2016). Distributions of sums of nonidentical random variables. In W. J. van der Linden (Hrsg.), *Handbook of item response theory* (S. 87–104). Boca Raton, FL: CRC Press.

Van der Linden, W. J. & Glas, C. A. W. (Hrsg.). (2010). *Elements of adaptive testing*. New York: Springer.

Van der Linden, W. J. & Pashley, P. P. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Hrsg.), *Elements of adaptive testing*. New York: Springer.

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/10.1177/01466216980223006

Van Deursen, A. & van Dijk, J. (2009). Using the Internet: Skill related problems in users' online behavior. *Interacting with Computers*, *21*, 393–402. https://doi.org/10.1016/j.intcom.2009.06.005

Van Deursen, A. & van Dijk, J. (2011). Internet skills and the digital divide. *New Media & Society*, *13*(6), 893–911. https://doi.org/10.1177/1461444810386774

Veerkamp, W. J. J. & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*, 203–226. https://doi.org/10.3102/10769986022002203

Veldkamp, B. P. (2010). Bayesian item selection in constrained adaptive testing using shadow tests. *Psicológica*, *31*, 149–169. Verfügbar unter https://www.researchgate.net

Veldkamp, B. P. & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*(4), 575–588. https://doi.org/10.1007/BF02295132

Wainer, H. (Hrsg.). (2000). *Computerized adaptive testing: A primer* (2. Aufl.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Wainer, H. & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Hrsg.), *Computerized adaptive testing: A primer* (2. Aufl., S. 61–100). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Walter, O. & Rost, J. (2011). 2. Kapitel: Psychometrische Grundlagen von Large Scale Assessments. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Methoden der psychologischen Diagnostik* (S. 88–150). Göttingen: Hogrefe.

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika*, *80*, 428–449. https://doi.org/10.1007/s11336-013-9399-0

Wang, C. & Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive testing—Gaining information from different angles. *Psychometrika*, *76*, 363–384. https://doi.org/10.1007/s11336-011-9215-7

Wang, C., Weiss, D. J. & Shang, Z. (2019). Variable-length stopping rules for multidimensional computerized adaptive testing. *Psychometrika*, *84*, 749–771. https://doi.org/10.1007/s11336-018-9644-7

Wang, W.-C. & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*, 295–316. https://doi.org/10.1177/0146621604265938

Wang, W.-C., Chen, P.-H. & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological methods*, *9*, 116–135. https://doi.org/10.1177/01466216980223001

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. Verfügbar unter https://link.springer.com

Weiss, D. J. & Guyer, R. (2012). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing*. St. Paul MN: Assessment Systems Corporation.

Wenzel, S. F. C., Engelhardt, L., Hartig, K., Kuchta, K., Goldhammer, F., Naumann, J. et al. (2016). Computergestützte, adaptive und verhaltensnahe Erfassung informations- und kommunikationstechnologiebezogener Fertigkeiten (ICT-Skills). In Referat Bildungsforschung (Hrsg.), *Forschungsvorhaben in Ankopplung an Large-Scale-Assessments* (S. 161–180). Berlin.

Wenzel, S. F. C., Engelhardt, L., Kuchta, K., Hartig, K., Naumann, J., Goldhammer, F. et al. (2015). Computergestützte, adaptive und verhaltensnahe Erfassung Informations-

und Kommunikations-technologie-bezogener Fähigkeiten (ICT-Skills) in PISA. Verfügbar unter http://edok01.tib.uni-hannover.de/edoks/e01fb15/838771823.pdf

Whittaker, S. & Sidner, C. (1996). E-mail overload: exploring personal information management of e-mail. In *Proceedings of the SIGCHI conference on Human factors in computing systems* .

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research Online*, (3), 1–22. Verfügbar unter www.mpr-online.de

Wise, S. L. & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica*, *21*, 135–155. Verfügbar unter https://www.redalyc.org/articulo.oa?id=16921108

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). ACER ConQuest Version 2: Generalised item response modelling [Computer software].

Yao, L. (2014). Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constraints. *Journal of Educational Measurement*, *51*, 18–38. Verfügbar unter https://link.springer.com

Yousfi, S. & Bohme, H. F. (2012). Principles and procedures of considering item sequence effects in the development of calibrated item pools. Conceptual analysis and empirical illustration. *Psychological Test and Assessment Modeling*, *54*, 366–396. Retrieved from http://www.wiso-net.de

Zabal, A., Martin, S., Klaukien, A., Rammstedt, B., Baumert, J. & Klieme, E. (2013). Grundlegende Kompetenzen der erwachsenen Bevölkerung in Deutschland im internationalen Vergleich. In B. Rammstedt (Hrsg.), *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich. Ergebnisse von PIAAC 2012* (S. 31–76). Münster: Waxmann.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal)*

Item Scores. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Anhangsverzeichnis

B.3 – Relative Messeffizienz

B.4 – Reliabilität

Anhang C: Tabellen zu Kapitel 7 – Studie II

Anhang D: Tabellen zu Kapitel 8 – Studie III

Anhang E: Erklärungen laut Promotionsordnung

Anhang F: Erklärung über die Eigenleistung

Anhang A: Tabelle und Abbildungen zu Kapitel 4

Tabelle A.1

Die 64 selektierten CavE-ICT-Items mit Zuordnung zu den im CavE-ICT-Framework definierten Facettenebenen, dargestellten Applikationen, Itemkennwerten und durchschnittlichen Bearbeitungszeiten

ID	Name	ICT-Framework:			Applikation	Mittlere Bearbeitungs-zeit	Item-schwierigkeit	*SE*	WMNSQ	*t*-Wert	Item-diskrimination	Antworten pro Item
		Kognitive Operation	Situation	Soziale Interaktion								
I001113	Translate	1	3	1	1	71.81	- 1.140	0.138	1.05	0.8	0.25	301
I004411	Theobromin	4	1	1	1	73.49	- 1.405	0.148	1.12	1.5	0.23	290
I005221	Badminton	2	1	2	2	98.55	- 1.410	0.140	1.05	0.6	0.26	321
I008513	Tortoise	5	3	1	5	58.59	- 0.047	0.133	1.00	0.1	0.41	256
I010122	ContactDetails	1	2	2	1	103.64	- 0.832	0.130	0.99	- 0.1	0.35	310
I011212	PrintDetails	2	2	1	4	89.22	2.562	0.203	1.02	0.2	0.17	311
I013413	Eyecolor	4	3	1	5	101.20	- 0.039	0.125	0.99	- 0.3	0.44	290
I015413	Download	4	3	1	1	40.91	- 0.452	0.134	1.04	0.9	0.29	261
I016213	DownloadFolder	2	3	1	3	70.23	- 0.064	0.126	1.03	0.7	0.45	289
I017322	YouthCommittee	3	2	2	3	133.17	2.156	0.178	1.03	0.3	0.14	316
I018513	CycleOfSeasons	5	3	1	4	47.02	- 2.362	0.199	0.98	- 0.1	0.29	300
I020211	E-MailFile	2	1	1	2	68.45	- 1.666	0.152	0.97	- 0.4	0.36	318
I023411	MailSorting	4	1	1	2	68.03	1.674	0.153	1.13	1.4	0.16	297
I024222	MRT	2	2	2	8	128.40	2.659	0.214	0.97	- 0.1	0.33	302
I025221	CarRace	2	1	2	8	85.30	- 0.813	0.126	1.00	0.1	0.40	319
I026411	Hotel	4	1	1	1	78.95	0.317	0.127	1.08	1.9	0.26	286
I026523	NameFolder	5	3	2	3	126.58	2.517	0.210	1.01	0.1	0.26	285
I029223	Shareboxx	2	3	2	1	158.73	- 0.183	0.132	1.05	1.1	0.31	265
I032512	MasterSlides	5	2	1	6	88.07	0.480	0.134	1.04	0.9	0.32	266
I033312	Shuttle	3	2	1	8	76.09	- 0.501	0.130	0.97	- 0.6	0.40	282
I034523	CoolChemist	5	3	2	1	82.04	- 0.130	0.123	0.99	- 0.1	0.39	306
I035211	ConvertDoc.	2	1	1	4	100.20	0.000	0.133	0.94	- 1.4	0.48	260
I037523	Wiki	5	3	2	1	129.73	0.463	0.133	0.92	- 1.8	0.52	264
I038222	SendMails	2	2	2	2	180.02	4.243	0.454	1.01	0.2	0.15	272
I039121	PicNeeded	1	1	2	1	134.13	- 0.829	0.134	1.04	0.8	0.35	287
I040113	LungCancer	1	3	1	1	125.36	1.913	0.165	1.02	0.3	0.27	306
I041321	Forum	3	1	2	1	107.04	0.020	0.133	1.04	0.9	0.43	258
I042512	EditPicture	5	2	1	8	52.70	- 2.568	0.212	1.03	0.2	0.22	275
I043322	Itinerary	3	2	2	1	241.04	2.736	0.238	0.93	- 0.3	0.26	271
I045211	InstallationChat	2	1	1	7	42.05	- 2.327	0.198	0.93	- 0.5	0.35	274

I047322	PlanningWork	3	2	2	2	176.93	3.541	0.325	1.01	0.1	0.11	290
I049212	CalcSpreadsheet	2	2	1	5	129.77	1.886	0.174	0.88	- 1.1	0.42	280
I050112	Sheets	1	2	1	5	109.60	2.060	0.182	1.05	0.4	0.17	275
I051111	DatFind	1	1	1	3	62.63	- 1.390	0.145	0.90	- 1.3	0.47	313
I053521	EditChat	5	1	2	1	67.99	- 2.961	0.236	1.03	0.2	0.16	324
I054113	FindBook	1	3	1	8	99.83	3.467	0.326	0.98	0.0	0.18	260
I056211	SavePassword	2	1	1	1	73.14	1.812	0.162	1.02	0.3	0.25	310
I058323	DateSummer	3	3	2	1	158.35	- 0.765	0.128	0.92	- 1.5	0.52	315
I062411	BrowserAddon	4	1	1	1	54.10	- 0.023	0.125	1.07	1.8	0.31	291
I063212	RecentDoc	2	2	1	3	113.67	- 0.078	0.119	0.84	- 4.3	0.56	325
I064313	BookACourse	3	3	1	1	100.78	0.463	0.120	1.06	1.5	0.25	320
I066111	FindFlight	1	1	1	1	62.29	1.878	0.171	1.02	0.2	0.20	277
I067121	FindMail	1	1	2	2	137.19	0.081	0.124	0.95	- 1.4	0.51	298
I068422	SelectProduct	4	2	2	1	149.49	1.528	0.158	0.97	- 0.4	0.33	279
I069413	SelectUpdate	4	3	1	7	76.86	- 0.697	0.135	1.03	0.6	0.31	271
I070223	SharedFolder	2	3	2	3	80.66	2.152	0.189	1.02	0.2	0.24	274
I071122	CompMistake	1	2	2	2	139.13	0.477	0.129	0.89	- 2.5	0.55	287
I075223	AdaptSpreadsheet	2	3	2	5	155.35	2.342	0.187	0.91	- 0.7	0.36	308
I076222	SelectMedium	2	2	2	2	147.91	- 0.075	0.123	0.98	- 0.6	0.40	301
I078223	PrintPresentation	2	3	2	6	95.60	0.229	0.131	1.05	1.2	0.36	265
I079113	HelpLogin	1	3	1	1	142.66	1.571	0.159	1.00	0.0	0.42	278
I082223	ShareInfo	2	3	2	1	99.98	0.832	0.135	1.03	0.5	0.38	285
I083321	ChooseFriend	3	1	2	1	86.07	1.012	0.140	0.90	- 1.6	0.49	281
I086223	NewUser	2	3	2	2	96.15	- 0.196	0.134	0.98	- 0.4	0.50	254
I087111	SearchBackup	1	1	1	1	137.94	0.959	0.134	1.02	0.4	0.31	294
I088522	OfficialMail	5	2	2	2	106.96	0.819	0.138	1.06	1.0	0.40	272
I094212	FindE-Mail	2	2	1	2	67.35	- 1.032	0.143	0.88	- 1.9	0.51	272
I095413	UseSpreadsheet	4	3	1	5	73.45	- 1.886	0.175	0.95	- 0.4	0.40	263
I096512	MakeTable	5	2	1	5	110.33	- 0.089	0.134	0.98	- 0.6	0.39	256
I097311	SportCourse	3	1	1	1	144.20	- 0.294	0.129	0.92	- 2.1	0.47	278
I106113v2	SearchWord	1	3	1	4	154.20	0.357	0.129	1.09	1.9	0.31	284
I108421	Billiard	4	1	2	1	84.27	- 2.043	0.170	1.01	0.1	0.32	312
I110323	WareIncome	3	3	2	2	214.55	0.752	0.140	0.94	- 1.1	0.42	258
I114412	FindJob	4	2	1	1	88.31	0.175	0.136	1.01	0.3	0.37	245

Anmerkungen. Kognitive Operationen: 1 = Zugreifen, 2 = Managen, 3 = Integrieren, 4 = Bewerten, 5 = Erzeugen; Situationen: 1 = Persönlich, 2 = Beruflich, 3 = Bildungsbezogen; Soziale Interaktion: 1 = Individuelle, 2 = Kollektiv; Applikationen: 1 = Browser, 2 = E-Mail, 3 = Ordnerstruktur, 4 = Text-, Präsentations-, Tabellenkalkulationssoftware, 5 = Sonstige.

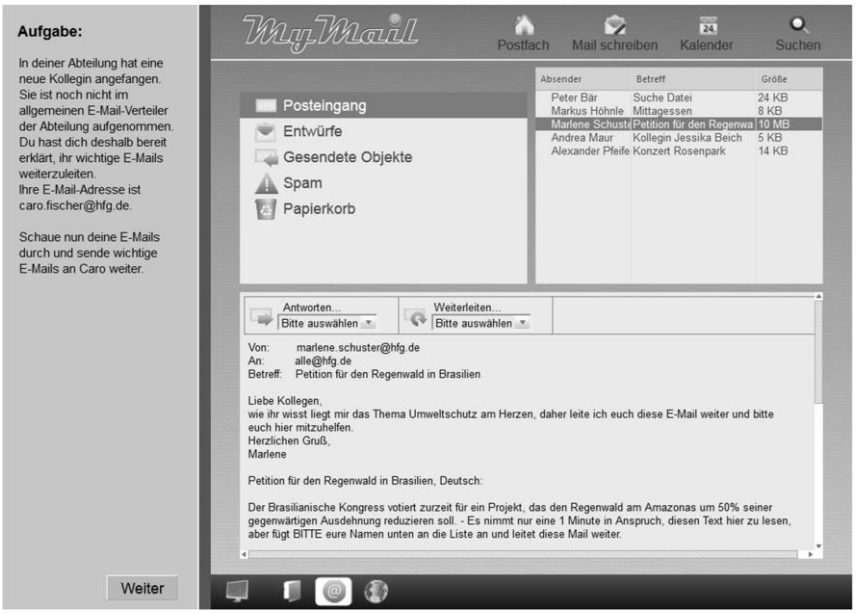

Abbildung A.1: Beispiel eines ICT-Items, welches den kognitiven Prozess „Bewerten“, in einer beruflichen Situation, mit einer kollektiven sozialen Interaktion abbildet (englischsprachige Version dieses Items in Engelhardt et al., 2017)

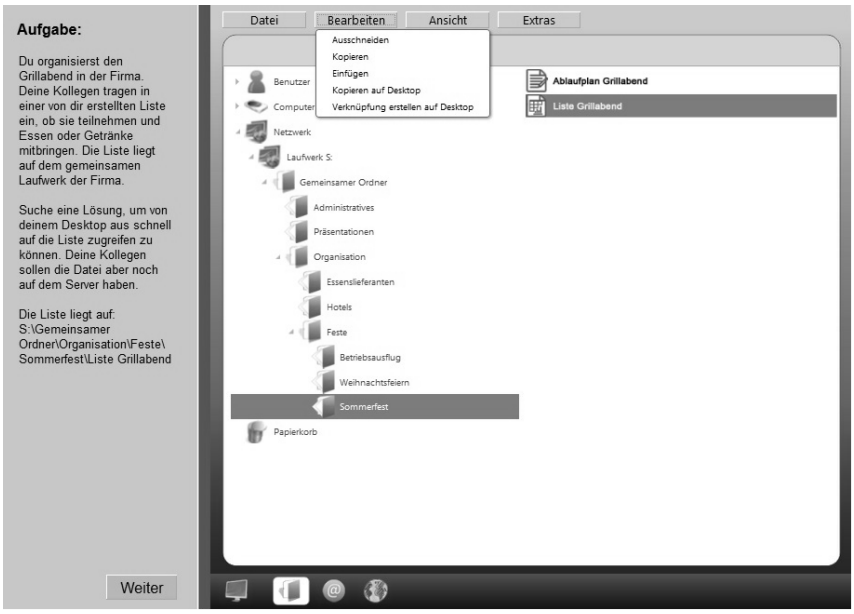

Abbildung A.2: Beispiel eines ICT-Items, welches den kognitiven Prozess „Managen“, in einer beruflichen Situation, mit einer kollektiven sozialen Interaktion abbildet

Anhang B: Tabellen zu Kapitel 6 – Studie I

B.1 - Bias

Tabelle B.1.1
Bias und Standardfehler (SE) für den MAT-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	MAT											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)
10	.50	50	.008	(.007)	- .013	(.007)	.011	(.007)	.001	(.006)	.001	(.007)	.001	(.007)
		75	.008	(.007)	- .013	(.006)	.011	(.006)	.000	(.006)	.002	(.006)	.001	(.006)
		100	.008	(.007)	- .013	(.007)	.011	(.007)	.001	(.007)	.002	(.007)	.002	(.007)
		125	.008	(.007)	- .015	(.007)	.008	(.006)	- .001	(.006)	.001	(.007)	.000	(.007)
	.70	50	- .016	(.006)	- .011	(.005)	.008	(.006)	.007	(.006)	- .003	(.005)	- .003	(.005)
		75	- .014	(.006)	- .009	(.006)	.010	(.006)	.009	(.005)	- .002	(.006)	- .001	(.006)
		100	- .016	(.007)	- .012	(.007)	.008	(.006)	.008	(.006)	- .004	(.006)	- .003	(.006)
		125	- .016	(.006)	- .011	(.005)	.008	(.005)	.007	(.005)	- .004	(.006)	- .003	(.005)
20	.50	50	.006	(.006)	- .011	(.007)	.010	(.007)	- .000	(.007)	.001	(.007)	.001	(.007)
		75	.007	(.006)	- .011	(.006)	.009	(.006)	.000	(.005)	.001	(.006)	.001	(.006)
		100	.006	(.008)	- .011	(.008)	.009	(.006)	.000	(.008)	.001	(.008)	.001	(.007)
		125	.007	(.007)	- .012	(.006)	.007	(.007)	- .001	(.007)	.001	(.007)	.000	(.007)
	.70	50	- .014	(.006)	- .009	(.005)	.009	(.006)	.008	(.006)	- .001	(.005)	- .001	(.006)
		75	- .012	(.005)	- .007	(.005)	.009	(.006)	.008	(.005)	- .002	(.006)	- .001	(.005)
		100	- .014	(.007)	- .011	(.007)	.008	(.006)	.007	(.006)	- .003	(.006)	- .003	(.006)
		125	- .013	(.006)	- .009	(.005)	.008	(.005)	.006	(.005)	- .003	(.005)	- .002	(.005)
40	.50	50	.005	(.009)	- .009	(.011)	.008	(.010)	- .003	(.013)	.002	(.010)	.001	(.011)
		75	.006	(.008)	- .008	(.006)	.007	(.006)	.001	(.007)	.001	(.007)	.001	(.007)
		100	.004	(.008)	- .010	(.007)	.006	(.006)	- .000	(.007)	- .000	(.009)	- .000	(.007)
		125	.006	(.007)	- .009	(.006)	.005	(.006)	- .001	(.006)	.001	(.006)	.000	(.006)
	.70	50	- .012	(.007)	- .007	(.008)	.009	(.008)	.006	(.010)	- .000	(.008)	- .001	(.008)
		75	- .009	(.006)	- .006	(.006)	.008	(.006)	.007	(.005)	- .001	(.006)	- .000	(.006)
		100	- .011	(.007)	- .008	(.006)	.007	(.006)	.007	(.006)	- .001	(.005)	- .001	(.006)
		125	- .010	(.005)	- .007	(.004)	.007	(.005)	.005	(.005)	- .002	(.006)	- .001	(.005)

Anmerkungen. MAT-Algorithmus = Multidimensional Adaptive Testing.

Tabelle B.1.2

Bias und Standardfehler (SE) für den C-MAT I-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT I											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)
10	.50	50	.009	(.007)	- .012	(.008)	.012	(.008)	.001	(.009)	.003	(.008)	.003	(.008)
		75	.009	(.007)	- .012	(.007)	.011	(.007)	.001	(.006)	.002	(.007)	.002	(.007)
		100	.009	(.008)	- .014	(.008)	.011	(.007)	.001	(.007)	.002	(.008)	.002	(.008)
		125	.009	(.007)	- .013	(.007)	.010	(.008)	- .000	(.007)	.001	(.007)	.001	(.007)
	.70	50	- .015	(.006)	- .010	(.007)	.009	(.007)	.008	(.007)	- .002	(.007)	- .002	(.007)
		75	- .014	(.007)	- .010	(.007)	.009	(.007)	.008	(.007)	- .002	(.007)	- .002	(.007)
		100	- .015	(.007)	- .011	(.007)	.009	(.006)	.008	(.006)	- .002	(.006)	- .002	(.006)
		125	- .015	(.007)	- .010	(.007)	.010	(.007)	.008	(.007)	- .002	(.008)	- .002	(.007)
20	.50	50	.006	(.007)	- .011	(.008)	.010	(.008)	- .001	(.009)	.002	(.008)	.001	(.008)
		75	.008	(.006)	- .010	(.006)	.009	(.006)	.001	(.006)	.001	(.006)	.002	(.006)
		100	.007	(.007)	- .012	(.007)	.009	(.006)	.000	(.007)	.001	(.008)	.001	(.007)
		125	.007	(.007)	- .012	(.006)	.008	(.006)	- .001	(.007)	.000	(.006)	.000	(.006)
	.70	50	- .013	(.006)	- .008	(.007)	.010	(.007)	.007	(.008)	- .001	(.006)	- .001	(.007)
		75	- .013	(.005)	- .008	(.006)	.008	(.005)	.007	(.006)	- .002	(.006)	- .001	(.006)
		100	- .013	(.007)	- .010	(.006)	.008	(.006)	.007	(.005)	- .002	(.005)	- .002	(.006)
		125	- .013	(.006)	- .009	(.006)	.009	(.006)	.007	(.006)	- .002	(.006)	- .002	(.006)
40	.50	50	.005	(.009)	- .009	(.011)	.008	(.010)	- .002	(.013)	.002	(.010)	.001	(.011)
		75	.006	(.008)	- .008	(.006)	.007	(.007)	.001	(.007)	.001	(.007)	.001	(.007)
		100	.004	(.007)	- .010	(.007)	.006	(.007)	.000	(.007)	.000	(.008)	- .000	(.007)
		125	.001	(.006)	- .009	(.006)	.005	(.007)	- .002	(.007)	.001	(.007)	.000	(.006)
	.70	50	- .012	(.007)	- .007	(.008)	.008	(.008)	.005	(.010)	- .000	(.007)	- .001	(.008)
		75	- .009	(.006)	- .006	(.006)	.007	(.006)	.007	(.006)	- .001	(.006)	- .000	(.006)
		100	- .011	(.006)	- .009	(.006)	.007	(.005)	.006	(.006)	- .002	(.006)	- .002	(.006)
		125	- .009	(.006)	- .007	(.005)	.007	(.006)	.005	(.006)	- .002	(.006)	- .001	(.006)

Anmerkungen. C-MAT I-Algorithmus = Constrained Multidimensional Adaptive Testing I.

Tabelle B.1.3

Bias und Standardfehler für den C-MAT II-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT II											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)
10	.50	50	.008	(.008)	- .014	(.007)	.012	(.007)	.001	(.006)	.001	(.006)	.002	(.007)
		75	.009	(.007)	- .012	(.007)	.011	(.007)	.001	(.007)	.001	(.006)	.002	(.007)
		100	.008	(.008)	- .015	(.007)	.009	(.007)	- .000	(.006)	.001	(.007)	.001	(.007)
		125	.008	(.006)	- .014	(.006)	.009	(.006)	.000	(.006)	.001	(.006)	.001	(.006)
	.70	50	- .016	(.007)	- .010	(.007)	.010	(.008)	.007	(.007)	- .003	(.007)	- .003	(.007)
		75	- .014	(.007)	- .009	(.006)	.010	(.007)	.009	(.007)	- .002	(.006)	- .001	(.006)
		100	- .015	(.006)	- .011	(.007)	.008	(.006)	.007	(.006)	- .003	(.007)	- .003	(.007)
		125	- .016	(.006)	- .011	(.006)	.008	(.006)	.007	(.006)	- .003	(.008)	- .003	(.007)
20	.50	50	.007	(.008)	- .011	(.008)	.010	(.008)	- .000	(.009)	.002	(.008)	.001	(.008)
		75	.007	(.007)	- .010	(.006)	.008	(.006)	.000	(.007)	.000	(.006)	.001	(.007)
		100	.006	(.007)	- .013	(.008)	.007	(.007)	- .001	(.005)	- .000	(.007)	- .000	(.007)
		125	.008	(.007)	- .010	(.006)	.008	(.005)	.000	(.007)	.002	(.006)	.002	(.006)
	.70	50	- .014	(.006)	- .008	(.006)	.010	(.007)	.007	(.006)	- .002	(.005)	- .001	(.006)
		75	- .011	(.006)	- .007	(.005)	.009	(.006)	.009	(.005)	- .001	(.006)	- .000	(.006)
		100	- .013	(.006)	- .010	(.006)	.009	(.005)	.008	(.005)	- .002	(.006)	- .002	(.005)
		125	- .013	(.006)	- .009	(.005)	.008	(.005)	.007	(.006)	- .002	(.006)	- .002	(.006)
40	.50	50	.006	(.009)	- .009	(.011)	.008	(.010)	- .003	(.013)	.003	(.010)	.001	(.011)
		75	.006	(.008)	- .008	(.006)	.007	(.007)	.001	(.007)	.001	(.008)	.001	(.007)
		100	.004	(.007)	- .010	(.008)	.006	(.006)	.000	(.007)	- .000	(.008)	.000	(.007)
		125	.006	(.006)	- .009	(.006)	.004	(.006)	- .002	(.007)	.001	(.006)	.000	(.006)
	.70	50	- .012	(.008)	- .007	(.008)	.009	(.008)	.005	(.010)	- .000	(.008)	- .001	(.008)
		75	- .009	(.006)	- .006	(.006)	.007	(.006)	.007	(.005)	- .001	(.006)	- .000	(.006)
		100	- .010	(.006)	- .008	(.006)	.007	(.005)	.007	(.006)	- .001	(.005)	- .001	(.006)
		125	- .011	(.006)	- .007	(.005)	.006	(.005)	.005	(.006)	- .002	(.006)	- .002	(.006)

Anmerkungen. C-MAT II-Algorithmus = Constrained Multidimensional Adaptive Testing II.

Tabelle B.1.4

Bias und Standardfehler für den C-MAT III-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT III											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)
10	.50	50	.008	(.008)	- .013	(.008)	.011	(.009)	.011	(.009)	.001	(.008)	.002	(.008)
		75	.010	(.007)	- .011	(.006)	.012	(.006)	.012	(.006)	.002	(.006)	.003	(.006)
		100	.009	(.008)	- .013	(.008)	.010	(.007)	.010	(.007)	.002	(.008)	.002	(.007)
		125	.008	(.007)	- .014	(.007)	.009	(.007)	.009	(.007)	.001	(.007)	.001	(.007)
	.70	50	- .015	(.007)	- .010	(.008)	.010	(.008)	.010	(.008)	- .002	(.007)	- .002	(.008)
		75	- .014	(.007)	- .009	(.007)	.010	(.006)	.010	(.006)	- .002	(.007)	- .001	(.007)
		100	- .015	(.006)	- .010	(.007)	.009	(.006)	.009	(.006)	- .002	(.006)	- .002	(.006)
		125	- .016	(.006)	- .011	(.006)	.008	(.006)	.008	(.006)	- .003	(.007)	- .003	(.006)
20	.50	50	.007	(.007)	- .012	(.008)	.010	(.008)	.010	(.008)	.001	(.008)	.001	(.008)
		75	.008	(.007)	- .010	(.006)	.010	(.006)	.010	(.006)	.001	(.006)	.002	(.006)
		100	.006	(.007)	- .012	(.008)	.008	(.006)	.008	(.006)	.001	(.007)	.000	(.007)
		125	.007	(.007)	- .013	(.005)	.006	(.007)	.006	(.007)	- .000	(.007)	- .001	(.007)
	.70	50	- .014	(.006)	- .009	(.006)	.009	(.007)	.009	(.007)	- .002	(.006)	- .002	(.006)
		75	- .011	(.006)	- .007	(.006)	.009	(.006)	.009	(.006)	- .002	(.007)	- .001	(.006)
		100	- .013	(.006)	- .010	(.006)	.008	(.006)	.008	(.006)	- .002	(.006)	- .002	(.006)
		125	- .013	(.006)	- .009	(.006)	.008	(.006)	.008	(.006)	- .002	(.006)	- .002	(.006)
40	.50	50	.006	(.009)	- .009	(.011)	.008	(.010)	.008	(.010)	.002	(.010)	.001	(.011)
		75	.006	(.008)	- .008	(.005)	.007	(.007)	.007	(.007)	.001	(.007)	.002	(.007)
		100	.004	(.007)	- .010	(.007)	.006	(.007)	.006	(.007)	.000	(.008)	.000	(.007)
		125	.006	(.006)	- .010	(.005)	.004	(.006)	.004	(.006)	- .000	(.006)	- .000	(.006)
	.70	50	- .012	(.007)	- .007	(.008)	.009	(.008)	.009	(.008)	- .000	(.008)	- .001	(.008)
		75	- .009	(.007)	- .006	(.006)	.007	(.006)	.007	(.006)	- .002	(.006)	- .000	(.006)
		100	- .010	(.006)	- .008	(.006)	.008	(.005)	.008	(.005)	- .001	(.006)	- .001	(.006)
		125	- .011	(.005)	- .007	(.005)	.006	(.005)	.006	(.005)	- .002	(.005)	- .002	(.005)

Anmerkungen. C-MAT III-Algorithmus = Constrained Multidimensional Adaptive Testing III.

Tabelle B.1.5

Bias und Standardfehler für den S-UAT-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	S-UAT											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)	Bias	(*SE*)
10	.50	50	.010	(.007)	- .009	(.009)	.012	(.008)	.001	(.011)	.003	(.010)	.003	(.009)
		75	.008	(.008)	- .010	(.007)	.011	(.008)	.000	(.007)	.001	(.008)	.002	(.008)
		100	.009	(.008)	- .012	(.007)	.010	(.006)	.000	(.007)	.003	(.008)	.002	(.007)
		125	.009	(.008)	- .010	(.006)	.009	(.008)	.001	(.007)	.003	(.007)	.002	(.007)
	.70	50	- .012	(.007)	- .009	(.009)	.006	(.008)	.004	(.010)	- .004	(.009)	- .003	(.008)
		75	- .013	(.009)	- .009	(.007)	.004	(.007)	.003	(.006)	- .004	(.008)	- .004	(.007)
		100	- .013	(.008)	- .011	(.008)	.005	(.007)	.003	(.007)	- .003	(.006)	- .004	(.007)
		125	- .014	(.008)	- .009	(.007)	.006	(.007)	.005	(.007)	- .003	(.007)	- .003	(.007)
20	.50	50	.005	(.008)	- .007	(.009)	.010	(.009)	- .001	(.011)	.003	(.007)	.002	(.009)
		75	.007	(.007)	- .005	(.008)	.008	(.007)	.001	(.007)	.002	(.006)	.003	(.007)
		100	.007	(.007)	- .008	(.006)	.008	(.006)	.001	(.007)	.003	(.008)	.002	(.007)
		125	.007	(.008)	- .009	(.007)	.008	(.007)	- .000	(.008)	.002	(.007)	.002	(.007)
	.70	50	- .011	(.008)	- .007	(.008)	.006	(.009)	.002	(.011)	- .003	(.008)	- .003	(.009)
		75	- .010	(.006)	- .004	(.008)	.001	(.007)	.002	(.007)	- .003	(.007)	- .003	(.007)
		100	- .011	(.009)	- .008	(.007)	.004	(.007)	.003	(.006)	- .002	(.007)	- .003	(.007)
		125	- .010	(.008)	- .007	(.007)	.005	(.007)	.002	(.007)	- .003	(.007)	- .003	(.007)
40	.50	50	.005	(.010)	- .005	(.013)	.006	(.012)	- .004	(.015)	.004	(.011)	.001	(.012)
		75	.005	(.009)	- .005	(.006)	.006	(.009)	.001	(.007)	.000	(.008)	.001	(.008)
		100	.003	(.007)	- .007	(.007)	.005	(.008)	.001	(.008)	.001	(.008)	.001	(.008)
		125	.005	(.007)	- .006	(.007)	.003	(.007)	- .001	(.007)	.002	(.007)	.001	(.007)
	.70	50	- .009	(.009)	- .005	(.012)	.003	(.013)	- .002	(.015)	.000	(.011)	- .002	(.012)
		75	- .008	(.010)	- .005	(.012)	.004	(.012)	- .000	(.014)	- .000	(.011)	- .002	(.012)
		100	- .008	(.008)	- .008	(.007)	.002	(.007)	.003	(.008)	- .002	(.008)	- .003	(.008)
		125	- .005	(.007)	- .005	(.006)	.002	(.006)	- .000	(.007)	- .002	(.007)	- .002	(.007)

Anmerkungen. S-UAT-Algorithmus = Sequential Unidimensional Adaptive Testing.

B.2 – Mean Squared Error

Tabelle B.2.1

Mean Squared Error (MSE) und Standardfehler (SE) für den MAT-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	MAT											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			MSE	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)
10	.50	50	.604	(.010)	.605	(.012)	.591	(.013)	.598	(.013)	.596	(.013)	.599	(.012)
		75	.603	(.010)	.603	(.009)	.588	(.008)	.597	(.010)	.593	(.010)	.597	(.010)
		100	.603	(.010)	.606	(.009)	.588	(.009)	.594	(.010)	.594	(.009)	.597	(.009)
		125	.602	(.009)	.604	(.009)	.587	(.007)	.597	(.009)	.595	(.008)	.597	(.008)
	.70	50	.509	(.010)	.497	(.009)	.514	(.010)	.505	(.013)	.496	(.009)	.504	(.010)
		75	.507	(.009)	.499	(.007)	.512	(.009)	.502	(.009)	.496	(.006)	.503	(.008)
		100	.506	(.009)	.496	(.007)	.508	(.009)	.501	(.009)	.497	(.008)	.502	(.009)
		125	.508	(.008)	.496	(.008)	.508	(.008)	.500	(.008)	.495	(.008)	.501	(.008)
20	.50	50	.456	(.018)	.458	(.018)	.450	(.015)	.458	(.022)	.449	(.018)	.454	(.018)
		75	.456	(.014)	.454	(.010)	.442	(.011)	.450	(.011)	.446	(.010)	.450	(.011)
		100	.452	(.010)	.455	(.010)	.441	(.008)	.447	(.007)	.445	(.010)	.448	(.009)
		125	.450	(.009)	.451	(.009)	.440	(.007)	.448	(.007)	.446	(.007)	.447	(.008)
	.70	50	.371	(.012)	.366	(.012)	.378	(.011)	.374	(.016)	.363	(.011)	.370	(.012)
		75	.371	(.009)	.364	(.007)	.373	(.008)	.366	(.009)	.360	(.006)	.367	(.008)
		100	.370	(.007)	.363	(.008)	.368	(.007)	.364	(.006)	.361	(.008)	.365	(.007)
		125	.370	(.006)	.362	(.006)	.369	(.006)	.366	(.006)	.360	(.005)	.365	(.006)
40	.50	50	.341	(.019)	.343	(.017)	.339	(.016)	.341	(.022)	.338	(.019)	.341	(.019)
		75	.322	(.018)	.325	(.016)	.314	(.013)	.321	(.017)	.318	(.015)	.320	(.016)
		100	.314	(.010)	.319	(.014)	.309	(.009)	.313	(.009)	.312	(.011)	.313	(.011)
		125	.311	(.008)	.312	(.009)	.305	(.008)	.312	(.009)	.309	(.008)	.310	(.008)
	.70	50	.281	(.014)	.276	(.012)	.287	(.012)	.280	(.016)	.274	(.012)	.289	(.013)
		75	.264	(.011)	.263	(.011)	.265	(.010)	.264	(.011)	.258	(.011)	.263	(.011)
		100	.260	(.007)	.256	(.009)	.258	(.007)	.256	(.007)	.254	(.008)	.257	(.008)
		125	.256	(.006)	.252	(.007)	.257	(.005)	.256	(.007)	.251	(.005)	.255	(.006)

Anmerkungen. MAT-Algorithmus = Multidimensional Adaptive Testing.

Tabelle B.2.2

Mean Squared Error (MSE) und Standardfehler (SE) für den C-MAT I-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT I											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			MSE	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)
10	.50	50	.627	(.008)	.626	(.012)	.612	(.009)	.617	(.013)	.618	(.011)	.620	(.011)
		75	.625	(.011)	.627	(.009)	.609	(.010)	.619	(.009)	.616	(.010)	.619	(.010)
		100	.623	(.009)	.624	(.009)	.607	(.010)	.618	(.010)	.616	(.007)	.617	(.009)
		125	.622	(.009)	.624	(.010)	.607	(.010)	.619	(.010)	.616	(.010)	.618	(.010)
	.70	50	.532	(.009)	.523	(.009)	.538	(.010)	.526	(.012)	.519	(.010)	.528	(.010)
		75	.532	(.011)	.523	(.009)	.535	(.008)	.527	(.008)	.520	(.009)	.527	(.009)
		100	.530	(.010)	.522	(.008)	.535	(.010)	.526	(.010)	.520	(.009)	.527	(.010)
		125	.531	(.009)	.520	(.007)	.532	(.009)	.525	(.010)	.519	(.008)	.525	(.009)
20	.50	50	.468	(.008)	.469	(.009)	.459	(.009)	.464	(.011)	.463	(.010)	.465	(.009)
		75	.467	(.009)	.468	(.008)	.455	(.009)	.463	(.007)	.460	(.009)	462	(.008)
		100	.464	(.009)	.466	(.008)	.454	(.009)	.461	(.007)	.458	(.007)	.461	(.008)
		125	.463	(.007)	.465	(.008)	.451	(.008)	.463	(.008)	.458	(.008)	.460	(.008)
	.70	50	.385	(.007)	.378	(.007)	.390	(.009)	.383	(.009)	.376	(.008)	.383	(.008)
		75	.383	(.009)	.378	(.007)	.387	(.007)	.381	(.007)	.375	(.006)	.381	(.007)
		100	.381	(.007)	.376	(.006)	.383	(.008)	.380	(.007)	.374	(.007)	.379	(.007)
		125	.383	(.006)	.375	(.006)	.383	(.007)	.378	(.007)	.372	(.007)	.378	(.007)
40	.50	50	.344	(.012)	.346	(.011)	.340	(.011)	.345	(.016)	.342	(.014)	.343	(.013)
		75	.328	(.010)	.328	(.008)	.320	(.008)	.325	(.008)	.323	(.007)	.325	(.008)
		100	.320	(.007)	.323	(.007)	.315	(.006)	.319	(.006)	.317	(.007)	.319	(.007)
		125	.318	(.006)	.319	(.007)	.312	(.005)	.317	(.006)	.315	(.006)	.316	(.006)
	.70	50	.284	(.009)	.279	(.009)	.289	(.009)	.284	(.011)	.278	(.009)	.283	(.009)
		75	.270	(.007)	.266	(.007)	.272	(.005)	.268	(.006)	.264	(.005)	.268	(.006)
		100	.266	(.005)	.261	(.005)	.265	(.005)	.264	(.005)	.260	(.005)	.263	(.005)
		125	.265	(.004)	.260	(.005)	.264	(.005)	.261	(.005)	.258	(.005)	.262	(.005)

Anmerkungen. C-MAT I-Algorithmus = Constrained Multidimensional Adaptive Testing I.

Tabelle B.2.3

Mean Squared Error (MSE) und Standardfehler (SE) für den C-MAT II-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT II											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			MSE	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)
10	.50	50	.621	(.010)	.607	(.009)	.591	(.009)	.598	(.009)	.596	(.009)	.603	(.009)
		75	.619	(.010)	.604	(.008)	.588	(.008)	.597	(.007)	.597	(.009)	.601	(.009)
		100	.618	(.010)	.605	(.007)	.588	(.008)	.598	(.008)	.594	(.008)	.600	(.008)
		125	.617	(.008)	.604	(.008)	.586	(.010)	.596	(.009)	.593	(.009)	.599	(.009)
	.70	50	.521	(.009)	.501	(.009)	.515	(.007)	.505	(.010)	.500	(.010)	.509	(.009)
		75	.519	(.008)	.500	(.008)	.513	(.008)	.505	(.008)	.499	(.006)	.507	(.008)
		100	.519	(.009)	.501	(.008)	.511	(.009)	.504	(.008)	.497	(.009)	.506	(.009)
		125	.520	(.009)	.499	(.008)	.510	(.010)	.502	(.008)	.496	(.008)	.506	(.008)
20	.50	50	.466	(.008)	.462	(.009)	.449	(.007)	.456	(.009)	.453	(.010)	.457	(.009)
		75	.466	(.010)	.456	(.009)	.445	(.008)	.452	(.008)	.450	(.007)	.454	(.008)
		100	.463	(.008)	.455	(.006)	.443	(.007)	.450	(.007)	.447	(.007)	.452	(.007)
		125	.463	(.008)	.454	(.007)	.442	(.008)	.450	(.007)	.446	(.009)	.451	(.008)
	.70	50	.382	(.007)	.371	(.006)	.381	(.008)	.375	(.008)	.367	(.007)	.375	(.007)
		75	.379	(.007)	.367	(.006)	.376	(.007)	.372	(.007)	.366	(.006)	.372	(.007)
		100	.379	(.007)	.367	(.007)	.374	(.006)	.369	(.007)	.363	(.006)	.370	(.006)
		125	.379	(.007)	.366	(.008)	.374	(.007)	.368	(.005)	.363	(.006)	.370	(.007)
40	.50	50	.344	(.012)	.346	(.011)	.339	(.011)	.344	(.016)	.341	(.014)	.343	(.013)
		75	.327	(.010)	.324	(.008)	.317	(.007)	.322	(.008)	.321	(.006)	.322	(.008)
		100	.321	(.007)	.319	(.007)	.311	(.006)	.316	(.005)	.313	(.006)	.316	(.006)
		125	.317	(.006)	.314	(.006)	.308	(.006)	.313	(.006)	.310	(.006)	.313	(.006)
	.70	50	.284	(.009)	.278	(.009)	.288	(.009)	.283	(.011)	.277	(.009)	.282	(.009)
		75	.269	(.007)	.263	(.007)	.269	(.006)	.266	(.006)	.262	(.005)	.266	(.006)
		100	.266	(.005)	.257	(.006)	.262	(.004)	.260	(.005)	.257	(.006)	.261	(.005)
		125	.264	(.005)	.256	(.005)	.260	(.004)	.258	(.004)	.254	(.005)	.258	(.004)

Anmerkungen. C-MAT II-Algorithmus = Constrained Multidimensional Adaptive Testing II.

Tabelle B.2.4

Mean Squared Error (MSE) und Standardfehler (SE) für den C-MAT III-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT III											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			MSE	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)
10	.50	50	.621	(.010)	.606	(.010)	.591	(.008)	.598	(.009)	.597	(.010)	.603	(.009)
		75	.620	(.010)	.603	(.007)	.587	(.009)	.598	(.009)	.598	(.009)	.601	(.009)
		100	.619	(.010)	.605	(.009)	.591	(.010)	.597	(.008)	596	(.010)	.602	(.009)
		125	.619	(.008)	.606	(.009)	.587	(.008)	.597	(.009)	.594	(.009)	.601	(.009)
	.70	50	.522	(.008)	.502	(.009)	.517	(.009)	.506	(.008)	.501	(.009)	.510	(.009)
		75	.520	(.009)	.503	(.008)	.515	(.008)	.507	(.007)	.500	(.007)	.509	(.008)
		100	.522	(.010)	.502	(.009)	.514	(.008)	.507	(.009)	.450	(.007)	.509	(.009)
		125	.522	(.009)	.501	(.008)	.513	(.008)	.505	(.008)	.499	(.008)	.508	(.008)
20	.50	50	.466	(.008)	.461	(.009)	.450	(.008)	.457	(.010)	.454	(.009)	.458	(.009)
		75	.466	(.009)	.456	(.006)	.445	(.008)	.453	(.007)	.451	(.007)	.454	(.008)
		100	.463	(.008)	.456	(.007)	.445	(.008)	.451	(.007)	.447	(.007)	.453	(.007)
		125	.463	(.008)	.455	(.007)	.443	(.007)	.450	(.008)	.448	(.007)	.452	(.007)
	.70	50	.382	(.007)	.371	(.007)	.382	(.007)	.377	(.007)	.368	(.007)	.376	(.007)
		75	.380	(.007)	.369	(.007)	.378	(.006)	.374	(.006)	.367	(.005)	.374	(.006)
		100	.380	(.008)	.368	(.006)	.376	(.006)	.372	(.007)	.365	(.006)	.372	(.006)
		125	.380	(.007)	.367	(.007)	.376	(.006)	.370	(.006)	.364	(.006)	.371	(.006)
40	.50	50	.344	(.012)	.345	(.011)	.339	(.011)	.344	(.016)	.341	(.014)	.343	(.013)
		75	.327	(.010)	.324	(.008)	.316	(.007)	.322	(.008)	.321	(.006)	.322	(.008)
		100	.321	(.007)	.319	(.007)	.311	(.006)	.315	(.006)	.313	(.007)	.316	(.007)
		125	.318	(.006)	.314	(.006)	.307	(.006)	.313	(.006)	.310	(.005)	.312	(.006)
	.70	50	.283	(.009)	.278	(.009)	.288	(.009)	.283	(.011)	.277	(.009)	.282	(.009)
		75	.269	(.007)	.263	(.007)	.269	(.005)	.266	(.006)	.262	(.005)	.266	(.006)
		100	.266	(.005)	.258	(.005)	.261	(.004)	.260	(.005)	.257	(.005)	.261	(.005)
		125	.264	(.005)	.257	(.005)	.261	(.004)	.258	(.004)	.254	(.004)	.259	(.005)

Anmerkungen. C-MAT IIII-Algorithmus = Constrained Multidimensional Adaptive Testing III.

Tabelle B.2.5

Mean Squared Error (MSE) und Standardfehler (SE) für den S-UAT-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	S-UAT											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			MSE	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)	*MSE*	(*SE*)
10	.50	50	.730	(.010)	.731	(.014)	.719	(.011)	.721	(.04)	.725	(.014)	.725	(.013)
		75	.729	(.013)	.732	(.010)	.715	(.012)	.722	(.012)	.724	(.012)	.724	(.012)
		100	.726	(.011)	.729	(.012)	.713	(.011)	.722	(.012)	.725	(.010)	.723	(.011)
		125	.726	(.010)	.729	(.010)	.713	(.010)	.722	(.013)	.723	(.010)	.723	(.011)
	.70	50	.721	(.010)	.711	(.012)	.730	(.013)	.716	(.016)	.712	(.012)	.718	(.013)
		75	.721	(.015)	.714	(.012)	.725	(.010)	.713	(.011)	.711	(.011)	.717	(.012)
		100	.720	(.014)	.713	(.010)	.726	(.011)	.714	(.011)	.711	(.011)	.717	(.011)
		125	.721	(.010)	.711	(.010)	.722	(.011)	.714	(.010)	.710	(.009)	.716	(.010)
20	.50	50	.558	(.011)	.561	(.011)	.553	(.010)	.557	(.014)	.557	(.012)	.557	(.011)
		75	.557	(.012)	.559	(.010)	.548	(.012)	.552	(.011)	.554	(.011)	.554	(.011)
		100	.554	(.010)	.557	(.010)	.547	(.010)	.552	(.008)	.551	(.008)	.552	(.009)
		125	.554	(.008)	.554	(.009)	.544	(.009)	.553	(.011)	.551	(.010)	.551	(.010)
	.70	50	.553	(.012)	.547	(.012)	.561	(.013)	.554	(.017)	.549	(.012)	.553	(.013)
		75	.550	(.013)	.546	(.011)	.555	(.010)	.547	(.010)	.546	(.010)	.549	(.011)
		100	.549	(.011)	.545	(.008)	.553	(.010)	.545	(.010)	.545	(.010)	.547	(.010)
		125	.550	(.011)	.544	(.009)	.552	(.009)	.545	(.009)	.543	(.009)	.547	(.009)
40	.50	50	.406	(.016)	.409	(.015)	.405	(.014)	.409	(.021)	.407	(.018)	.407	(.017)
		75	.385	(.013)	.386	(.010)	.380	(.010)	.382	(.011)	.383	(.009)	.383	(.011)
		100	.375	(.008)	.379	(.009)	.373	(.008)	.375	(.008)	.376	(.009)	.376	(.008)
		125	.373	(.007)	.373	(.008)	.369	(.008)	.371	(.008)	.372	(.007)	.372	(.007)
	.70	50	.404	(.018)	.401	(.018)	.413	(.016)	.407	(.022)	.401	(.017)	.405	(.018)
		75	.381	(.013)	.378	(.010)	.383	(.010)	.380	(.010)	.380	(.011)	.380	(.011)
		100	.374	(.009)	.371	(.007)	.375	(.008)	.371	(.008)	.372	(.009)	.373	(.008)
		125	.372	(.008)	.370	(.009)	.372	(.007)	.368	(.007)	.367	(.008)	.370	(.008)

Anmerkungen. S-UAT-Algorithmus = Sequential Unidimensional Adaptive Testing.

B.3 – Relative Messeffizienz

Tabelle B.3.1
Relative Messeffizienz (RE) und Standardfehler (SE) für den MAT-Algorithmus im Vergleich zum S-UAT-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	MAT											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			RE	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)
10	.50	50	1.207	(.022)	1.209	(.023)	1.218	(.024)	1.207	(.022)	1.218	(.029)	1.212	(.024)
		75	1.209	(.019)	1.213	(.019)	1.217	(.024)	1.209	(.022)	1.220	(.023)	1.214	(.021)
		100	1.205	(.023)	1.205	(.022)	1.212	(.020)	1.215	(.022)	1.220	(.023)	1.211	(.022)
		125	1.206	(.019)	1.206	(.022)	1.215	(.019)	1.210	(.028)	1.214	(.021)	1.210	(.022)
	.70	50	1.416	(.025)	1.431	(.025)	1.421	(.029)	1.417	(.030)	1.434	(.027)	1.424	(.027)
		75	1.422	(.033)	1.432	(.027)	1.416	(.026)	1.419	(.033)	1.434	(.030)	1.425	(.030)
		100	1.421	(.029)	1.437	(.023)	1.428	(.028)	1.426	(.028)	1.433	(.026)	1.429	(.027)
		125	1.420	(.028)	1.434	(.028)	1.422	(.032)	1.427	(.030)	1.435	(.031)	1.428	(.030)
20	.50	50	1.225	(.037)	1.225	(.038)	1.230	(.031)	1.216	(.034)	1.244	(.036)	1.228	(.035)
		75	1.221	(.026)	1.232	(.020)	1.239	(.026)	1.227	(.028)	1.242	(.024)	1.232	(.025)
		100	1.226	(.024)	1.225	(.024)	1.241	(.026)	1.235	(.020)	1.238	(.027)	1.233	(.024)
		125	1.230	(.025)	1.229	(.024)	1.238	(.022)	1.235	(.024)	1.237	(.023)	1.234	(.024)
	.70	50	1.491	(.036)	1.496	(.033)	1.486	(.036)	1.484	(.033)	1.512	(.035)	1.494	(.035)
		75	1.484	(.029)	1.501	(.027)	1.489	(.029)	1.495	(.031)	1.516	(.029)	1.497	(.029)
		100	1.485	(.028)	1.500	(.034)	1.502	(.031)	1.496	(.030)	1.511	(.033)	1.499	(.031)
		125	1.486	(.031)	1.505	(.029)	1.495	(.033)	1.490	(.030)	1.508	(.030)	1.497	(.030)
40	.50	50	1.190	(.027)	1.193	(.026)	1.197	(.025)	1.201	(.027)	1.206	(.026)	1.198	(.026)
		75	1.195	(.034)	1.188	(.037)	1.214	(.031)	1.191	(.037)	1.206	(.035)	1.199	(.035)
		100	1.197	(.028)	1.191	(.037)	1.208	(.021)	1.197	(.028)	1.205	(.023)	1.200	(.027)
		125	1.201	(.026)	1.197	(.023)	1.211	(.025)	1.193	(.024)	1.207	(.028)	1.202	(.025)
	.70	50	1.436	(.024)	1.452	(.022)	1.437	(.028)	1.451	(.023)	1.461	(.027)	1.447	(.025)
		75	1.444	(.028)	1.441	(.033)	1.444	(.040)	1.438	(.036)	1.473	(.038)	1.448	(.035)
		100	1.442	(.033)	1.453	(.042)	1.454	(.026)	1.450	(.033)	1.464	(.034)	1.453	(.034)
		125	1.453	(.029)	1.468	(.030)	1.447	(.029)	1.438	(.028)	1.462	(.035)	1.454	(.030)

Anmerkungen. MAT-Algorithmus = Multidimensional Adaptive Testing.

Tabelle B.3.2

Relative Messeffizienz (RE) und Standardfehler (SE) für den C-MAT I-Algorithmus im Vergleich zum S-UAT-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT I											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			RE	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)
10	.50	50	1.164	(.011)	1.168	(.012)	1.176	(.012)	1.169	(.009)	1.174	(.010)	1.170	(.011)
		75	1.166	(.010)	1.168	(.011)	1.174	(.010)	1.166	(.011)	1.175	(.012)	1.170	(.011)
		100	1.167	(.012)	1.169	(.010)	1.175	(.011)	1.169	(.011)	1.177	(.009)	1.171	(.011)
		125	1.167	(.008)	1.169	(.011)	1.174	(.010)	1.167	(.014)	1.174	(.011)	1.170	(.011)
	.70	50	1.355	(.019)	1.360	(.019)	1.357	(.020)	1.360	(.017)	1.371	(.020)	1.361	(.019)
		75	1.355	(.019)	1.366	(.016)	1.356	(.016)	1.352	(.016)	1.369	(.017)	1.359	(.017)
		100	1.356	(.018)	1.365	(.016)	1.357	(.019)	1.358	(.019)	1.369	(.018)	1.361	(.018)
		125	1.358	(.018)	1.367	(.015)	1.358	(.018)	1.361	(.017)	1.370	(.017)	1.363	(.017)
20	.50	50	1.192	(.012)	1.196	(.010)	1.205	(.011)	1.199	(.012)	1.205	(.013)	1.199	(.012)
		75	1.192	(.010)	1.195	(.012)	1.205	(.010)	1.191	(.012)	1.205	(.011)	1.198	(.011)
		100	1.194	(.011)	1.195	(.010)	1.205	(.011)	1.194	(.012)	1.204	(.010)	1.198	(.011)
		125	1.195	(.011)	1.192	(.011)	1.205	(.011)	1.194	(.011)	1.204	(.011)	1.198	(.011)
	.70	50	1.435	(.022)	1.447	(.021)	1.439	(.023)	1.447	(.020)	1.458	(.025)	1.445	(.022)
		75	1.435	(.021)	1.447	(.017)	1.436	(.017)	1.436	(.019)	1.457	(.020)	1.442	(.019)
		100	1.439	(.019)	1.449	(.017)	1.442	(.019)	1.436	(.020)	1.457	(.018)	1.445	(.018)
		125	1.434	(.021)	1.453	(.018)	1.439	(.020)	1.442	(.023)	1.458	(.019)	1.445	(.020)
40	.50	50	1.178	(.014)	1.182	(.011)	1.193	(.014)	1.186	(.013)	1.192	(.013)	1.186	(.013)
		75	1.173	(.011)	1.177	(.011)	1.187	(.010)	1.174	(.010)	1.185	(.010)	1.179	(.011)
		100	1.174	(.010)	1.174	(.010)	1.183	(.011)	1.176	(.010)	1.185	(.011)	1.178	(.011)
		125	1.174	(.009)	1.171	(.011)	1.185	(.010)	1.172	(.011)	1.183	(.009)	1.177	(.010)
	.70	50	1.423	(.028)	1.437	(.028)	1.428	(.027)	1.433	(.027)	1.443	(.027)	1.433	(.028)
		75	1.411	(.023)	1.425	(.020)	1.409	(.018)	1.414	(.021)	1.439	(.020)	1.419	(.020)
		100	1.409	(.022)	1.422	(.016)	1.414	(.019)	1.408	(.019)	1.431	(.021)	1.417	(.019)
		125	1.407	(.018)	1.424	(.019)	1.410	(.019)	1.409	(.017)	1.425	(.019)	1.415	(.019)

Anmerkungen. C-MAT I-Algorithmus = Constrained Multidimensional Adaptive Testing I.

Tabelle B.3.3

Relative Messeffizienz (RE) und Standardfehler (SE) für den C-MAT II-Algorithmus im Vergleich zum S-UAT-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT II											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			RE	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)
10	.50	50	1.176	(.016)	1.205	(.025)	1.218	(.023)	1.206	(.025)	1.217	(.024)	1.204	(.023)
		75	1.177	(.016)	1.211	(.020)	1.217	(.025)	1.209	(.024)	1.211	(.024)	1.205	(.022)
		100	1.176	(.020)	1.206	(.023)	1.213	(.026)	1.208	(.022)	1.221	(.024)	1.205	(.023)
		125	1.176	(.019)	1.207	(.022)	1.217	(.023)	1.212	(.024)	1.218	(.021)	1.206	(.022)
	.70	50	1.385	(.023)	1.419	(.029)	1.417	(.028)	1.417	(.029)	1.423	(.026)	1.412	(.027)
		75	1.389	(.031)	1.429	(.033)	1.413	(.028)	1.412	(.027)	1.425	(.027)	1.414	(.029)
		100	1.387	(.027)	1.425	(.031)	1.421	(.026)	1.415	(.023)	1.432	(.031)	1.416	(.027)
		125	1.388	(.029)	1.424	(.024)	1.416	(.035)	1.421	(.030)	1.431	(.028)	1.416	(.029)
20	.50	50	1.197	(.017)	1.214	(.020)	1.231	(.019)	1.220	(.017)	1.230	(.021)	1.218	(.019)
		75	1.196	(.018)	1.227	(.021)	1.230	(.022)	1.220	(.023)	1.230	(.024)	1.221	(.022)
		100	1.195	(.022)	1.225	(.021)	1.233	(.018)	1.226	(.020)	1.233	(.022)	1.223	(.021)
		125	1.196	(.022)	1.220	(.025)	1.231	(.026)	1.229	(.023)	1.235	(.027)	1.222	(.024)
	.70	50	1.446	(.025)	1.476	(.026)	1.474	(.030)	1.477	(.025)	1.494	(.029)	1.473	(.027)
		75	1.452	(.029)	1.489	(.027)	1.477	(.029)	1.471	(.029)	1.494	(.035)	1.477	(.030)
		100	1.449	(.035)	1.486	(.032)	1.478	(.027)	1.478	(.029)	1.501	(.033)	1.478	(.031)
		125	1.450	(.029)	1.489	(.030)	1.474	(.035)	1.479	(.028)	1.497	(.031)	1.478	(.031)
40	.50	50	1.178	(.014)	1.185	(.012)	1.197	(.014)	1.190	(.014)	1.195	(.014)	1.189	(.014)
		75	1.176	(.016)	1.190	(.014)	1.201	(.016)	1.185	(.015)	1.194	(.016)	1.189	(.015)
		100	1.171	(.019)	1.190	(.018)	1.198	(.019)	1.187	(.020)	1.199	(.016)	1.189	(.018)
		125	1.176	(.023)	1.188	(.020)	1.200	(.022)	1.185	(.018)	1.199	(.017)	1.190	(.020)
	.70	50	1.124	(.028)	1.442	(.029)	1.432	(.028)	1.437	(.029)	1.448	(.028)	1.437	(.028)
		75	1.415	(.028)	1.441	(.023)	1.425	(.024)	1.427	(.027)	1.452	(.024)	1.432	(.025)
		100	1.405	(.025)	1.442	(.024)	1.431	(.022)	1.427	(.021)	1.448	(.025)	1.431	(.023)
		125	1.413	(.028)	1.447	(.028)	1.430	(.027)	1.430	(.027)	1.446	(.029)	1.433	(.028)

Anmerkungen. C-MAT II-Algorithmus = Constrained Multidimensional Adaptive Testing II.

Tabelle B.3.4

Relative Messeffizienz (RE) und Standardfehler (SE) für den C-MAT III-Algorithmus im Vergleich zum S-UAT-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT III											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			RE	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)	*RE*	(*SE*)
10	.50	50	1.176	(.017)	1.205	(.025)	1.217	(.022)	1.207	(.025)	1.215	(.026)	1.204	(.023)
		75	1.177	(.019)	1.214	(.020)	1.218	(.025)	1.208	(.023)	1.211	(.026)	1.205	(.023)
		100	1.174	(.020)	1.206	(.027)	1.206	(.023)	1.210	(.022)	1.216	(.025)	1.202	(.023)
		125	1.174	(.017)	1.203	(.025)	1.214	(.024)	1.208	(.025)	1.216	(.023)	1.203	(.023)
	.70	50	1.382	(.021)	1.418	(.028)	1.413	(.031)	1.414	(.030)	1.421	(.027)	1.409	(.028)
		75	1.385	(.032)	1.421	(.030)	1.407	(.030)	1.407	(.028)	1.422	(.028)	1.409	(.029)
		100	1.378	(.029)	1.421	(.028)	1.412	(.026)	1.409	(.028)	1.424	(.026)	1.409	(.027)
		125	1.382	(.026)	1.418	(.025)	1.407	(.032)	1.413	(.033)	1.425	(.027)	1.409	(.029)
20	.50	50	1.198	(.017)	1.215	(.018)	1.229	(.018)	1.219	(.018)	1.229	(.019)	1.218	(.018)
		75	1.195	(.017)	1.227	(.016)	1.232	(.023)	1.218	(.022)	1.227	(.020)	1.220	(.020)
		100	1.195	(.022)	1.221	(.026)	1.229	(.020)	1.224	(.020)	1.232	(.021)	1.220	(.022)
		125	1.197	(.022)	1.217	(.022)	1.227	(.024)	1.228	(.025)	1.232	(.024)	1.220	(.024)
	.70	50	1.446	(.026)	1.475	(.028)	1.468	(.030)	1.472	(.031)	1.491	(.027)	1.470	(.028)
		75	1.448	(.031)	1.480	(.030)	1.467	(.028)	1.465	(.026)	1.490	(.028)	1.470	(.029)
		100	1.444	(.036)	1.482	(.028)	1.472	(.029)	1.467	(.030)	1.492	(.032)	1.472	(.031)
		125	1.446	(.031)	1.482	(.031)	1.468	(.034)	1.474	(.028)	1.491	(.028)	1.472	(.030)
40	.50	50	1.178	(.014)	1.185	(.012)	1.196	(.04)	1.190	(.014)	1.195	(.014)	1.189	(.014)
		75	1.176	(.016)	1.191	(.013)	1.202	(.015)	1.186	(.015)	1.194	(.014)	1.189	(.015)
		100	1.171	(.018)	1.190	(.018)	1.198	(.017)	1.190	(.017)	1.201	(.014)	1.190	(.017)
		125	1.176	(.023)	1.187	(.016)	1.202	(.019)	1.188	(.018)	1.199	(.015)	1.190	(.018)
	.70	50	1.424	(.028)	1.442	(.029)	1.432	(.028)	1.437	(.029)	1.448	(.028)	1.437	(.028)
		75	1.414	(.290)	1.439	(.022)	1.422	(.023)	1.426	(.026)	1.452	(.024)	1.431	(.025)
		100	1.406	(.027)	1.438	(.022)	1.434	(.021)	1.426	(.024)	1.448	(.025)	1.431	(.024)
		125	1.414	(.028)	1.442	(.025)	1.429	(.029)	1.426	(.026)	1.447	(.026)	1.432	(.027)

Anmerkungen. C-MAT III-Algorithmus = Constrained Multidimensional Adaptive Testing III.

B.4 – Reliabilität

Tabelle B.4.1
Reliabilität (REL) und Standardfehler (SE) für den MAT-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	MAT											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			REL	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)
10	.50	50	.395	(.012)	.395	(.015)	.393	(.015)	.391	(.016)	.398	(.015)	.394	(.015)
		75	.396	(.012)	.396	(.011)	.396	(.010)	.392	(.012)	.400	(.012)	.396	(.012)
		100	.396	(.012)	.394	(.011)	.396	(.011)	.395	(.012)	.399	(.011)	.396	(.021)
		125	.397	(.012)	.395	(.011)	.397	(.009)	.392	(.011)	.398	(.010)	.396	(.010)
	.70	50	.483	(.013)	.485	(.012)	.483	(.013)	.479	(.018)	.487	(.013)	.483	(.014)
		75	.485	(.013)	.484	(.010)	.485	(.012)	.482	(.013)	.487	(.008)	.485	(.011)
		100	.486	(.012)	.486	(.010)	.488	(.012)	.484	(.012)	.487	(.011)	.486	(.012)
		125	.484	(.010)	.487	(.011)	.489	(.010)	.484	(.011)	.489	(.011)	.486	(.011)
20	.50	50	.543	(.026)	.542	(.026)	.538	(.022)	.533	(.031)	.546	(.026)	.541	(.026)
		75	.543	(.019)	.546	(.015)	.546	(.016)	.542	(.016)	.549	(.015)	.545	(.016)
		100	.548	(.014)	.545	(.014)	.547	(.012)	.544	(.010)	.550	(.014)	.547	(.013)
		125	.550	(.013)	.549	(.013)	.548	(.010)	.544	(.01)	.549	(.010)	.548	(.012)
	.70	50	.623	(.020)	.621	(.020)	.620	(.018)	.615	(.026)	.625	(.019)	.621	(.020)
		75	.624	(.016)	.623	(.012)	.625	(.014)	.622	(.014)	.628	(.010)	.624	(.013)
		100	.625	(.011)	.624	(.014)	.630	(.012)	.624	(.011)	.627	(.014)	.626	(.012)
		125	.624	(.010)	6.62	(.011)	.628	(.010)	.623	(.011)	.628	(.009)	.626	(.010)
40	.50	50	.659	(.033)	.657	(.030)	.652	(.029)	.653	(.039)	.659	(.035)	.656	(.033)
		75	.678	(.032)	.675	(.030)	.678	(.025)	.673	(.031)	.679	(.028)	.677	(.029)
		100	.686	(.020)	.681	(.026)	.683	(.018)	.681	(.017)	.685	(.021)	.683	(.020)
		125	.689	(.015)	.688	(.017)	.687	(.016)	.683	(.017)	.688	(.016)	.687	(.016)
	.70	50	.715	(.028)	.714	(.025)	.711	(.025)	.711	(.032)	.717	(.026)	.714	(.028)
		75	.732	(.024)	.728	(.024)	.733	(.022)	.728	(.023)	.734	(.025)	.731	(.024)
		100	.736	(.016)	.735	(.020)	.741	(.015)	.736	(.017)	.738	(.019)	.737	(.017)
		125	.740	(.013)	.739	(.015)	.741	(.012)	.736	(.015)	.741	(.013)	.739	(.014)

Anmerkungen. MAT-Algorithmus = Multidimensional Adaptive Testing.

Tabelle B.4.2

Reliabilität (REL) und Standardfehler (SE) für den C-MAT I-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT I											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			REL	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)
10	.50	50	.373	(.010)	.374	(.014)	.371	(.011)	.372	(.015)	.375	(.013)	.373	(.013)
		75	.374	(.013)	.373	(.011)	.374	(.012)	.369	(.011)	.378	(.012)	.374	(.012)
		100	.377	(.011)	.376	(.010)	.377	(.012)	.371	(.012)	.377	(.009)	.376	(.011)
		125	.377	(.011)	.376	(.012)	.376	(.012)	.370	(.013)	.377	(.013)	.375	(.012)
	.70	50	.460	(.012)	.459	(.012)	.459	(.013)	.457	(.016)	.464	(.013)	.460	(.014)
		75	.460	(.014)	.459	(.012)	.462	(.011)	.456	(.011)	.463	(.013)	.460	(.012)
		100	.461	(.014)	.460	(.011)	.462	(.013)	.458	(.014)	.464	(.012)	.461	(.013)
		125	.461	(.012)	.462	(.010)	.465	(.012)	.459	(.014)	.464	(.010)	.462	(.012)
20	.50	50	.531	(.011)	.531	(.013)	.529	(.012)	.527	(.016)	.532	(.015)	.530	(.013)
		75	.533	(.013)	.532	(.011)	.533	(.013)	.529	(.011)	.535	(.013)	.532	(.012)
		100	.536	(.012)	.534	(.011)	.534	(.014)	.530	(.010)	.538	(.010)	.534	(.011)
		125	.537	(.010)	.535	(.012)	.536	(.011)	.528	(.011)	.537	(.012)	.535	(.011)
	.70	50	.609	(.012)	.609	(.011)	.608	(.014)	.605	(.015)	.611	(.013)	.609	(.013)
		75	.611	(.015)	.609	(.012)	.611	(.011)	.607	(.012)	.613	(.011)	.610	(.012)
		100	.613	(.012)	.611	(.010)	.615	(.012)	.609	(.012)	.614	(.012)	.612	(.012)
		125	.611	(.011)	.612	(.011)	.615	(.012)	.611	(.011)	.616	(.011)	.613	(.011)
40	.50	50	.656	(.021)	.654	(.020)	.651	(.020)	.649	(.027)	.655	(.025)	.653	(.022)
		75	.672	(.018)	.672	(.014)	.671	(.014)	.669	(.015)	.673	(.013)	.672	(.015)
		100	.680	(.013)	.677	(.012)	.676	(.012)	.675	(.012)	.679	(.013)	.678	(.013)
		125	.682	(.011)	.681	(.013)	.680	(.010)	.677	(.011)	.682	(.011)	.680	(.011)
	.70	50	.712	(.019)	.711	(.019)	.710	(.018)	.708	(.023)	.713	(.019)	.711	(.020)
		75	.726	(.015)	.725	(.015)	.727	(.012)	.724	(.013)	.727	(.012)	.726	(.041)
		100	.730	(.011)	.730	(.011)	.734	(.011)	.728	(.012)	.732	(.012)	.731	(.011)
		125	.731	(.010)	.731	(.011)	.734	(.011)	.731	(.011)	.734	(.012)	.732	(.011)

Anmerkungen. C-MAT I-Algorithmus = Constrained Multidimensional Adaptive Testing I.

Tabelle B.4.3

Reliabilität (REL) und Standardfehler (SE) für den C-MAT II-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT II											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			REL	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)
10	.50	50	.379	(.012)	.393	(.011)	.393	(.011)	.391	(.011)	.397	(.011)	.391	(.011)
		75	.380	(.012)	.396	(.010)	.396	(.010)	.392	(.009)	.396	(.011)	.392	(.010)
		100	.382	(.012)	.395	(.009)	.396	(.010	.391	(.010)	.400	(.010)	.393	(.010)
		125	.382	(.009)	.396	(.009)	.398	(.012)	.393	(.011)	.400	(.011)	.394	(.011)
	.70	50	.472	(.012)	.481	(.012)	.482	(.010)	.479	(.014)	.483	(.013)	.479	(.012)
		75	.473	(.011)	.483	(.011)	.484	(.011)	.480	(.010)	.485	(.008)	.481	(.011)
		100	.473	(.012)	.482	(.011)	.486	(.012)	.480	(.011)	.487	(.012)	.482	(.021)
		125	.473	(.012)	.483	(.010)	.487	(.013)	.482	(.010)	.487	(.011)	.482	(.011)
20	.50	50	.534	(.012)	.538	(.012)	.539	(.011)	.535	(.013)	.542	(.014)	.537	(.012)
		75	.534	(.014)	.545	(.013)	.542	(.011)	.540	(.012)	.545	(.011)	.541	(.012)
		100	.537	(.011)	.546	(.009)	.544	(.010)	.541	(.011)	.548	(.010)	.543	(.010)
		125	.537	(.011)	.546	(.011)	.546	(.012)	.542	(.010)	.549	(.013)	.544	(.011)
	.70	50	.612	(.011)	.616	(.010)	.617	(.013)	.613	(.013)	.621	(.012)	.616	(.012)
		75	.620	(.012)	.622	(.011)	.622	(.011)	.616	(.012)	.622	(.010)	.619	(.011)
		100	.616	(.012)	.621	(.011)	.624	(.010)	.620	(.011)	.625	(.010)	.621	(.011)
		125	.615	(.012)	.622	(.013)	.624	(.011)	.620	(.008)	.626	(.011)	.621	(.011)
40	.50	50	.656	(.021)	.655	(.020)	.652	(.019)	.650	(.027)	.656	(.025)	.654	(.022)
		75	.673	(.018)	.676	(.015)	.675	(.014)	.672	(.015)	.676	(.012)	.674	(.015)
		100	.679	(.013)	.681	(.014)	.680	(.012)	.678	(.010)	.683	(.012)	.680	(.012)
		125	.682	(.012)	.686	(.011)	.684	(.013)	.681	(.012)	.686	(.012)	.684	(.012)
	.70	50	.713	(.019)	.712	(.019)	.711	(.018)	.709	(.023)	.715	(.019)	.712	(.019)
		75	.727	(.015)	.728	(.015)	.730	(.012)	.726	(.014)	.730	(.012)	.728	(.014)
		100	.730	(.011)	.734	(.012)	.737	(.010)	.732	(.011)	.735	(.012)	.733	(.011)
		125	.733	(.011)	.735	(.011)	.738	(.010)	.735	(.009)	.738	(.010)	.736	(.010)

Anmerkungen. C-MAT II-Algorithmus = Constrained Multidimensional Adaptive Testing II.

Tabelle B.4.4

Reliabilität (REL) und Standardfehler (SE) für den C-MAT III-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	C-MAT III											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			REL	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)
10	.50	50	.379	(.012)	.394	(.012)	.393	(.009)	.392	(.011)	.396	(.012)	.391	(.011)
		75	.380	(.013)	.397	(.009)	.397	(.012)	.391	(.011)	.396	(.012)	.392	(.013)
		100	.381	(.013)	.395	(.010)	.393	(.013)	.392	(.009)	.398	(.012)	.392	(.011)
		125	.381	(.009)	.395	(.011)	.397	(.010)	.392	(.011)	.400	(.012)	.393	(.011)
	.70	50	.471	(.011)	.481	(.012)	.481	(.012)	.478	(.011)	.483	(.012)	.479	(.012)
		75	.472	(.012)	.480	(.011)	.482	(.010)	.478	(.010)	.484	(.010)	.479	(.011)
		100	.470	(.013)	.481	(.012)	.483	(.011)	.478	(.012)	.484	(.010)	.479	(.012)
		125	.470	(.011)	.481	(.011)	.484	(.011)	.479	(.011)	.485	(.011)	.480	(.011)
20	.50	50	.534	(.011)	.539	(.013)	.538	(.011)	.535	(.015)	.541	(.014)	.537	(.013)
		75	.534	(.013)	.545	(.009)	.543	(.012)	.539	(.011)	.544	(.011)	.541	(.011)
		100	.537	(.012)	.544	(.010)	.543	(.012)	.541	(.010)	.548	(.010)	.542	(.011)
		125	.537	(.011)	.545	(.010)	.545	(.011)	.542	(.012)	.548	(.011)	.543	(.011)
	.70	50	.612	(.012)	.616	(.012)	.616	(.012)	.612	(.012)	.620	(.012)	.615	(.012)
		75	.615	(.012)	.618	(.013)	.620	(.010)	.615	(.011)	.622	(.008)	.618	(.011)
		100	.615	(.013)	.620	(.010)	.623	(.010)	.617	(.011)	.623	(.011)	.620	(.011)
		125	.614	(.012)	.620	(.012)	.622	(.010)	.619	(.010)	.624	(.011)	.620	(.011)
40	.50	50	.656	(.021)	.655	(.020)	.652	(.020)	.650	(.027)	.656	(.025)	.654	(.022)
		75	.673	(.018)	.676	(.014)	.675	(.013)	.672	(.015)	.676	(.012)	.674	(.014)
		100	.679	(.014)	.682	(.013)	.680	(.011)	.679	(.011)	.684	(.013)	.681	(.012)
		125	.682	(.012)	.686	(.011)	.684	(.012)	.682	(.012)	.686	(.009)	.684	(.011)
	.70	50	.713	(.019)	.712	(.019)	.711	(.018)	.709	(.023)	.715	(.019)	.712	(.019)
		75	.727	(.015)	.728	(.015)	.729	(.012)	.726	(.013)	.730	(.011)	.728	(.013)
		100	.730	(.011)	.733	(.012)	.737	(.010)	.732	(.011)	.735	(.012)	.733	(.011)
		125	.733	(.011)	.734	(.012)	.738	(.010)	.734	(.010)	.738	(.010)	.735	(.011)

Anmerkungen. C-MAT III-Algorithmus = Constrained Multidimensional Adaptive Testing III.

Tabelle B.4.5

Reliabilität (REL) und Standardfehler (SE) für den S-UAT-Algorithmus differenziert für die fünf latenten Merkmalsdimensionen und über die Dimensionen gemittelt

Testlänge	Korrelation	Itempoolgröße	S-UAT											
			Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5		Gemittelt	
			REL	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)	*REL*	(*SE*)
10	.50	50	.269	(.011)	.268	(.015)	.260	(.012)	.265	(.016)	.266	(.016)	.266	(.014)
		75	.270	(.014)	.268	(.011)	.265	(.013)	.265	(.013)	.268	(.013)	.267	(.013)
		100	.273	(.012)	.270	(.013)	.267	(.012)	.265	(.013)	.267	(.011)	.268	(.012)
		125	.273	(.011)	.271	(.011)	.267	(.012)	.265	(.014)	.269	(.012)	.269	(.012)
	.70	50	.268	(.012)	.263	(.014)	.265	(.014)	.262	(.018)	.265	(.013)	.265	(.014)
		75	.268	(.017)	.260	(.013)	.270	(.011)	.265	(.012)	.265	(.013)	.266	(.013)
		100	.269	(.015)	.262	(.011)	.269	(.012)	.264	(.012)	.265	(.012)	.266	(.012)
		125	.267	(.011)	.264	(.011)	.273	(.012)	.264	(.012)	.266	(.010)	.267	(.011)
20	.50	50	.441	(.014)	.439	(.013)	.432	(.013)	.433	(.017)	.436	(.015)	.436	(.014)
		75	.443	(.014)	.441	(.012)	.437	(.015)	.438	(.014)	.440	(.014)	.440	(.014)
		100	.446	(.013)	.443	(.013)	.438	(.013)	.437	(.011)	.442	(.010)	.441	(.012)
		125	.446	(.010)	.445	(.012)	.441	(.012)	.437	(.014)	.442	(.013)	.442	(.012)
	.70	50	.438	(.015)	.433	(.015)	.435	(.016)	.428	(.021)	.433	(.015)	.434	(.017)
		75	.442	(.016)	.434	(.014)	.441	(.013)	.435	(.013)	.436	(.013)	.438	(.014)
		100	.443	(.014)	.436	(.010)	.444	(.013)	.438	(.013)	.437	(.013)	.439	(.012)
		125	.442	(.014)	.436	(.011)	.445	(.012)	.438	(.012)	.439	(.012)	.440	(.012)
40	.50	50	.594	(.025)	.591	(.024)	.584	(.022)	.584	(.032)	.589	(.028)	.588	(.026)
		75	.615	(.021)	.614	(.016)	.609	(.016)	.611	(.018)	.613	(.015)	.613	(.017)
		100	.624	(.014)	.621	(.015)	.617	(.013)	.618	(.013)	.620	(.015)	.620	(.014)
		125	.626	(.012)	.627	(.013)	.621	(.013)	.622	(.013)	.624	(.011)	.624	(.012)
	.70	50	.590	(.028)	.585	(.028)	.585	(.025)	.581	(.034)	.587	(.027)	.586	(.028)
		75	.613	(.021)	.608	(.017)	.615	(.016)	.609	(.017)	.608	(.018)	.611	(.015)
		100	.620	(.015)	.616	(.012)	.623	(.013)	.617	(.013)	.616	(.014)	.618	(.014)
		125	.622	(.013)	.617	(.014)	.625	(.012)	.620	(.012)	.621	(.013)	.621	(.013)

Anmerkungen. S-UAT-Algorithmus = Sequential Unidimensional Adaptive Testing.

Anhang C: Tabellen zu Kapitel 7 – Studie II

Tabelle C.1

Relative Messeffizienz (RE) und Standardfehler (SE) für die multidimensionalen adaptiven Test-Algorithmen im Vergleich zum S-UAT-Algorithmus (Sequential Unidimensional Adaptive Testing) für die fünf latenten Merkmalsdimensionen gemittelt über Replikationen.

Testlänge	Dimension	Testalgorithmen: MAT *RE* (*SE*)	CU-MAT *RE* (*SE*)	C-MAT I *RE* (*SE*)	C-MAT III *RE* (*SE*)
10	1	1.211 (.049)	1.224 (.056)	1.196 (.036)	1.223 (.054)
	2	1.242 (.061)	1.296 (.061)	1.252 (.047)	1.303 (.059)
	3	1.237 (.053)	1.247 (.057)	1.196 (.032)	1.252 (.049)
	4	1.196 (.054)	1.187 (.055)	1.168 (.031)	1.193 (.060)
	5	1.220 (.050)	1.212 (.056)	1.171 (.032)	1.220 (.046)
20	1	1.334 (.048)	1.337 (.057)	1.312 (.038)	1.330 (.053)
	2	1.403 (.067)	1.468 (.071)	1.432 (.055)	1.468 (.069)
	3	1.315 (.052)	1.330 (.052)	1.295 (.033)	1.329 (.046)
	4	1.291 (.056)	1.270 (.060)	1.252 (.032)	1.269 (.050)
	5	1.276 (.043)	1.271 (.054)	1.246 (.030)	1.277 (.041)
30	1	1.361 (.050)	1.361 (.059)	1.305 (.034)	1.349 (.051)
	2	1.535 (.073)	1.528 (.074)	1.499 (.053)	1.530 (.072)
	3	1.313 (.045)	1.345 (.045)	1.322 (.034)	1.345 (.038)
	4	1.302 (.043)	1.279 (.048)	1.260 (.026)	1.277 (.040)
	5	1.274 (.037)	1.285 (.035)	1.266 (.029)	1.285 (.035)
40	1	1.379 (.044)	1.362 (.049)	1.343 (.034)	1.349 (.043)
	2	1.624 (.064)	1.524 (.066)	1.498 (.048)	1.516 (.057)
	3	1.319 (.039)	1.339 (.037)	1.330 (.034)	1.336 (.035)
	4	1.272 (.035)	1.257 (.039)	1.252 (.025)	1.258 (.031)
	5	1.247 (.033)	1.277 (.031)	1.273 (.030)	1.275 (.030)

Anmerkungen. MAT = Multidimensional Adaptive Testing; CU-MAT = Combined Unidimensional Multidimensional Adaptive Testing; C-MAT I & III = Constrained Multidimensional Adaptive Testing I & III; Dimensionen = Kognitive Prozesse: 1 = Zugreifen, 2 = Managen, 3 = Integrieren, 4 = Bewerten, 5 = Erzeugen.

Tabelle C.2

Reliabilität (REL) und Standardfehler (SE) für alle Test-Algorithmen für jede der fünf Merkmalsdimensionen gemittelt über Replikationen

Testlänge	Dimension	Testalgorithmen:				
		MAT	CU-MAT	C-MAT I	C-MAT III	S-UAT
		REL (*SE*)	*REL* (*SE*)	*REL* (*SE*)	*REL* (*SE*)	*REL* (*SE*)
10	1	.295 (.023)	.303 (.022)	.281 (.026)	.293 (.021)	.149 (.021)
	2	.325 (.024)	.360 (.025)	.334 (.027)	.359 (.022)	.168 (.021)
	3	.345 (.023)	.265 (.023)	.323 (.022)	.351 (.024)	.197 (.020)
	4	.282 (.023)	.315 (.024)	.247 (.024)	.264 (.026)	.133 (.020)
	5	.320 (.023)	.315 (.024)	.289 (.019)	.318 (.023)	.177 (.019)
20	1	.433 (.022)	.435 (.023)	.423 (.023)	.428 (.020)	.270 (.023)
	2	.488 (.024)	.512 (.026)	.499 (.027)	.510 (.022)	.307 (.025)
	3	.496 (.021)	.501 (.023)	.487 (.025)	.500 (.024)	.351 (.025)
	4	.400 (.025)	.382 (.024)	.372 (.025)	.381 (.026)	.243 (.024)
	5	.457 (.025)	.453 (.025)	.441 (.026)	.454 (.025)	.322 (.025)
30	1	.510 (.023)	.509 (.024)	.502 (.022)	.503 (.021)	.359 (.024)
	2	.596 (.023)	.593 (.023)	.585 (.024)	.593 (.022)	.404 (.024)
	3	.570 (.023)	.580 (.022)	.572 (.023)	.579 (.022)	.449 (.025)
	4	.463 (.027)	.452 (.027)	.444 (.026)	.451 (.027)	.323 (.026)
	5	.526 (.027)	.530 (.025)	.522 (.026)	.529 (.027)	.412 (.025)
40	1	.562 (.024)	.555 (.025)	.549 (.023)	.550 (.022)	.417 (.023)
	2	.663 (.023)	.641 (.025)	.635 (.023)	.639 (.021)	.476 (.025)
	3	.613 (.022)	.619 (.024)	.617 (.024)	.618 (.024)	.503 (.026)
	4	.500 (.028)	.493 (.028)	.491 (.029)	.492 (.028)	.381 (.029)
	5	.549 (.026)	.561 (.024)	.560 (.026)	.560 (.026)	.456 (.026)

Anmerkungen. MAT = Multidimensional Adaptive Testing; CU-MAT = Combined Unidimensional Multidimensional Adaptive Testing; C-MAT I & III = Constrained Multidimensional Adaptive Testing I & III; S-UAT = Sequential Unidimensional Adaptive Testing; Dimensionen = Kognitive Prozesse: 1 = Zugreifen, 2 = Managen, 3 = Integrieren, 4 = Bewerten, 5 = Erzeugen.

Anhang D: Tabellen zu Kapitel 8 – Studie III

Tabelle D.1

Items und Itemeigenschaften der vier manuell zusammengestellten ICT-Kurztests

Testlänge	ID	Name	ICT-Framework: Kognitive Operation	ICT-Framework: Situation	ICT-Framework: Soziale Interaktion	Applikation	Mittlere Bearbeitungszeit	Itemschwierigkeit	*SE*	WMNSQ	*t*-Wert	Itemdiskrimination
	I034523	CoolChemist	5	3	2	1	82.04	- 0.130	0.123	0.99	- 0.1	0.39
	I037523	Wiki	5	3	2	1	129.73	0.463	0.133	0.92	- 1.8	0.52
	I049212	CalcSpreadsheet	2	2	1	4	129.77	1.886	0.174	0.88	- 1.1	0.42
	I051111	DatFind	1	1	1	3	62.63	- 1.390	0.145	0.90	- 1.3	0.47
	I067121	FindMail	1	1	2	2	137.19	0.081	0.124	0.95	- 1.4	0.51
	I068422	SelectProduct	4	2	2	1	149.49	1.528	0.158	0.97	- 0.4	0.33
	I083321	ChooseFriend	3	1	2	1	86.07	1.012	0.140	0.90	- 1.6	0.49
	I094212	FindE-Mail	2	2	1	2	67.35	- 1.032	0.143	0.88	- 1.9	0.51
	I095413	UseSpreadsheet	4	3	1	4	73.45	- 1.885	0.175	0.95	- 0.4	0.40
10	I097311	SportCourse	3	1	1	5	144.20	- 0.294	0.129	0.92	- 2.1	0.47
	I032512	MasterSlides	5	2	1	4	88.07	0.480	0.134	1.04	0.9	0.32
	I058323	DateSummer	3	3	2	5	158.35	- 0.765	0.128	0.92	- 1.5	0.52
	I069413	SelectUpdate	4	3	1	5	76.86	- 0.697	0.135	1.03	0.6	0.31
	I086223	NewUser	2	3	2	5	96.15	- 0.196	0.134	0.98	- 0.4	0.50
15	I087111	SearchBackup	1	1	1	1	137.94	0.959	0.134	1.02	0.4	0.31
	I008513	Tortoise	5	3	1	4	58.59	- 0.047	0.133	1.00	0.1	0.41
	I033312	Shuttle	3	2	1	5	76.09	- 0.501	0.130	0.97	- 0.6	0.40
	I071122	CompMistake	1	2	2	3	139.13	0.477	0.129	0.89	- 2.5	0.55
	I082223	ShareInfo	2	3	2	1	99.98	0.832	0.135	1.03	0.5	0.38
20	I114412	FindJob	5	2	1	1	88.31	0.175	0.136	1.01	0.3	0.37
	I013413	Eyecolor	4	3	1	4	101.20	- 0.039	0.125	0.99	- 0.3	0.44
	I063212	RecentDoc	2	2	1	3	113.67	- 0.078	0.119	0.84	- 4.3	0.56
	I088522	OfficialMail	5	2	2	2	106.96	0.819	0.138	1.06	1.0	0.40
	I106113v2	SearchWord	1	3	1	1	154.20	0.357	0.129	1.09	1.9	0.31
25	I110323	WareIncome	3	3	2	4	214.55	0.752	0.140	0.94	- 1.1	0.42

Anmerkungen. Kognitive Operationen: 1 = Zugreifen, 2 = Managen, 3 = Integrieren, 4 = Bewerten, 5 = Erzeugen; Situationen: 1 = Persönlich, 2 = Beruflich, 3 = Bildungsbezogen; Soziale Interaktion: 1 = Individuelle, 2 = Kollektiv; Applikationen: 1 = Browser, 2 = E-Mail, 3 = Ordnerstruktur, 4 = Text-, Präsentations-, Tabellenkalkulationssoftware, 5 = Sonstige.

Tabelle D.2

Items und Itemeigenschaften der vier automatisiert zusammengestellten ICT-Kurztests

Testlänge	ID	Name	ICT-Framework: Kognitive Operation	ICT-Framework: Situation	ICT-Framework: Soziale Interaktion	Applikation	Mittlere Bearbeitungszeit	Itemschwierigkeit	*SE*	WMNSQ	*t*-Wert	Itemdiskrimination
	I032512	MasterSlides	5	2	1	4	88.07	0.480	0.134	1.04	0.9	0.32
	I037523	Wiki	5	3	2	1	129.73	0.463	0.133	0.92	- 1.8	0.52
	I041321	Forum	3	1	2	1	107.04	0.020	0.133	1.04	0.9	0.43
	I062411	BrowserAddon	4	1	1	1	54.10	- 0.023	0.125	1.07	1.8	0.31
	I067121	FindMail	1	1	2	2	137.19	0.081	0.124	0.95	- 1.4	0.51
	I071122	CompMistake	1	2	2	3	139.13	0.477	0.129	0.89	- 2.5	0.55
	I078223	PrintPresent	2	3	2	4	95.60	0.229	0.131	1.05	1.2	0.36
	I083321	ChooseFriend	3	1	2	1	86.07	1.012	0.140	0.90	- 1.6	0.49
	I086223	NewUser	2	3	2	5	96.15	- 0.196	0.134	0.98	- 0.4	0.50
10	I114412	FindJob	5	2	1	1	88.31	0.175	0.136	1.01	0.3	0.37
	I008513	Tortoise	5	3	1	4	58.59	- 0.047	0.133	1.00	0.1	0.41
	I013413	Eyecolor	4	3	1	4	101.20	- 0.039	0.125	0.99	- 0.3	0.44
	I035211	ConvertDoc	2	1	1	4	100.20	0.000	0.133	0.94	- 1.4	0.48
	I106113v2	SearchWord	1	3	1	1	154.20	0.357	0.129	1.09	1.9	0.31
15	I110323	WareIncome	3	3	2	4	214.55	0.752	0.140	0.94	- 1.1	0.42
	I016213	DownloadFolder	2	3	1	3	70.23	- 0.064	0.126	1.03	0.7	0.45
	I068422	SelectProduct	4	2	2	1	149.49	1.528	0.158	0.97	- 0.4	0.33
	I087111	SearchBackup	1	1	1	1	137.94	0.959	0.134	1.02	0.4	0.31
	I096512	MakeTable	5	2	1	4	110.33	- 0.089	0.134	0.98	- 0.6	0.39
20	I097311	SportCourse	3	1	1	5	144.20	- 0.294	0.129	0.92	- 2.1	0.47
	I033312	Shuttle	3	2	1	5	76.09	- 0.501	0.130	0.97	- 0.6	0.40
	I039121	PicNeeded	1	1	2	1	134.13	- 0.829	0.134	1.04	0.8	0.35
	I069413	SelectUpdate	4	3	1	5	76.86	- 0.697	0.135	1.03	0.6	0.31
	I082223	ShareInfo	2	3	2	1	99.98	0.832	0.135	1.03	0.5	0.38
25	I088522	OfficialMail	5	2	2	2	106.96	0.819	0.138	1.06	1.0	0.40

Anmerkungen. Kognitive Operationen: 1 = Zugreifen, 2 = Managen, 3 = Integrieren, 4 = Bewerten, 5 = Erzeugen; Situationen: 1 = Persönlich, 2 = Beruflich, 3 = Bildungsbezogen; Soziale Interaktion: 1 = Individuelle, 2 = Kollektiv; Applikationen: 1 = Browser, 2 = E-Mail, 3 = Ordnerstruktur, 4 = Text-, Präsentations-, Tabellenkalkulationssoftware, 5 = Sonstige.

Anhang E: Erklärungen laut Promotionsordnung

Erklärung zur Promotionsordnung

Ich erkläre hiermit, dass mir die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fachbereiche der Goethe-Universität Frankfurt am Main bekannt ist.

Frankfurt am Main, den 13.05.2020

S. Franziska C. Wenzel

Eidesstattliche Versicherung

Ich erkläre hiermit, dass ich die vorgelegte Dissertation mit dem Titel *Messung komplexer Konstrukte mit computerisierten Verfahren am Beispiel von ICT-Skills* selbständig angefertigt und mich anderer Hilfsmittel als der in ihr angegebenen nicht bedient habe, insbesondere, dass alle Entlehnungen aus anderen Schriften mit Angabe der betreffenden Schrift gekennzeichnet sind. Ich versichere, die Grundsätze der guten wissenschaftlichen Praxis beachtet, und nicht die Hilfe einer kommerziellen Promotionsvermittlung in Anspruch genommen zu haben.

Frankfurt am Main, den 13.05.2020

S. Franziska C. Wenzel

Erklärungen über frühere Promotionsversuche

Ich erkläre hiermit, dass ich mich bisher keiner Doktorprüfung unterzogen habe.

Frankfurt am Main, den 13.05.2020

S. Franziska C. Wenzel

Anhang F: Erklärung über die Eigenleistung

Erklärung über die Eigenleistung der Kandidatin

„Die Dissertation muss eine Erklärung enthalten, in der die Eigenleistung des Kandidaten/der Kandidatin dargestellt wird. Insbesondere bei Schriften mit Koautoren, aber auch bei in Einzelautorenschaft entstandenen Schriften, die oft auch im Rahmen von Abteilungsprojekten, Drittmittelprojekten, Projektverbünden usw. entstanden sind, soll dargelegt werden, welchen Anteil die Kandidaten an Entwicklung der Fragestellung, Design, Durchführung, Auswertung der empirischen Studie(n) und an dem Abfassen der einzelnen Beiträge hatten. Die Erklärung ist von dem/der BetreuerIn zu bestätigen."

Die vorliegende Dissertation mit dem Titel *Messung komplexer Konstrukte mit computerisierten Verfahren am Beispiel von ICT-Skills* umfasst drei empirische Studien, die auf Konstruktannahmen und Daten basieren die im Verbundprojekt CavE-ICT PISA erarbeitet wurden. Das Projekt wurde vom Bundesministerium für Bildung und Forschung (BMBF) gefördert und von Prof. Dr. Holger Horz (Goethe-Universität Frankfurt), Prof. Dr. Andreas Frey (Goethe-Universität Frankfurt), Prof. Dr. Frank Goldhammer (DIPF - Frankfurt) und Prof. Dr. Johannes Naumann (Bergische Universität Wuppertal) eingeworben. Im Rahmen des Projekts (Laufzeit 4/2012 - 3/2015) sollte ein Test zur computergestützten, verhaltensnahen und potentiell adaptiven Erfassung von Informations- und Kommunikationstechnologiebezogenen Fertigkeiten und Fähigkeiten (ICT-Skills) entwickelt und erprobt werden. Die Daten des CavE-ICT-Feldtests die unter anderem der IRT-Kalibrierung der im Projekt entwickelten ICT-Items dienten, wurden am DIPF ausgelesen und standen seit Herbst 2014 zur Verfügung.

Die Kandidatin arbeitete in dem Projekt an der Friedrich-Schiller-Universität Jena als wissenschaftliche Mitarbeiterin unter der Projektleitung von Prof. Dr. Andreas Frey im Teilprojekt *Psychometrie und adaptives Testen*.

Im Rahmen der vorliegenden Monografie erfolgt in Kapitel 4 eine Darstellung des Projekts und der Projektergebnisse, da diese die Grundlage der drei empirischen Studien dieser Dissertation bilden. Es wird dabei auf Veröffentlichungen die im Rahmen des Projekts entstanden sind verwiesen.

Ziel der drei Studien dieser Dissertation war es weitere, als die im Projekt fokussierten Testszenarien (Vorgabe der 64 Items umfassenden Gesamt-Skala und Einsatz eines eindimensionalen CAT) zum Einsatz der entwickelten ICT-Items mit dem Ziel der Messung von ICT-Skills bei möglichst umfassender Abbildung des Konstrukts zu explorieren.

Studie I: Spezifikation von Itempoolgröße und Testlänge eines computerisierten adaptiven Tests zur multidimensionalen Messung von ICT-Skills

Die übergeordnete Fragestellung sowie die Umsetzung der adaptiven Algorithmen im Rahmen dieser Simulationsstudie wurden von der Kandidatin unter Beratung von Prof. Dr. Andreas Frey entwickelt und umgesetzt. Die spezifische Ausformulierung der Fragestellungen, die Datengenerierung sowie die statistischen Analysen zur Auswertung der Simulationsstudie wurden von der Kandidatin eigenständig durchgeführt. Vorläufige Ergebnisse der Simulationsstudie wurden im Projektteam vorgestellt und diskutiert.

Studie II: Erprobung verschiedener CAT-Algorithmen unter Nutzung der CavE-ICT-Feldtestdaten

Die Fragestellungen wurden von der Kandidatin selbstständig entwickelt. Verschiedene adaptive Algorithmen, unter Nutzung von Daten des CavE-ICT-Feldtests, wurden von der Kandidatin im Rahmen einer Simulationsstudie erprobt. Die Datengenerierung sowie die statistischen Analysen zur Auswertung der Simulationsstudie wurden von der Kandidatin eigenständig durchgeführt.

Studie III: Zusammenstellung von Kurztests zur eindimensionalen Erfassung von ICT-Skills unter Nutzung der im Zuge des CavE-ICT-Feldtests geschätzten Itemparameter

Im Projektteam wurde die Nützlichkeit der Ableitung eines ICT-Kurztests diskutiert. Unter Beratung von Prof. Dr. Andreas Frey definierte die Kandidatin Kriterien und Möglichkeiten der Testzusammenstellungen. Die manuellen wie auch die automatisierten Testzusammenstellungen wurden von der Kandidatin eigenständig vorgenommen. Die Datengenerierung unter Nutzung von Daten des CavE-ICT-Feldtests sowie die statistischen Analysen zur Auswertung der Simulationsstudie wurden von der Kandidatin eigenständig durchgeführt.

Prof. Dr. Andreas Frey	S. Franziska C. Wenzel
(Betreuer der Dissertation)	(Verfasserin der Dissertation)