Der Übergang von papierbasiertem zu computerbasiertem Testen in Large-Scale Assessments

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

vorgelegt beim Fachbereich 05
der Johann Wolfgang Goethe - Universität

in Frankfurt am Main

von
Sarah Suzanne Bürger

aus Worms

Frankfurt 2017 (D 30)

vom Fachbereich 05 der

Johann Wolfgang Goethe - Universität

als Dissertation angenommen.

Dekan: Prof. Dr. Dr. Winfried Banzer

Gutachter: Prof. Dr. Frank Goldhammer

Prof. Dr. Johannes Hartig

Datum der Disputation: 25.10.2017

Inhaltsverzeichnis

Zusammenfassung

Die vorliegende Dissertation befasst sich mit dem Umstieg von papierbasiertem (PBA) auf computerbasiertes Assessment (CBA), insbesondere in Large-Scale-Studien. In der Bildungsforschung war Papier lange Zeit das Medium für Assessments, im Zuge des digitalen Zeitalters erhält der Computer aber auch hier Einzug. So sind die großen Bildungsvergleichsstudien, wie PISA (Programme for International Student Assessment) oder PIAAC (Programme for the International Assessment of Adult Competencies), und nationalen Studien über Bildungsverläufe und -entwicklungen im Rahmen des NEPS (Nationales Bildungspanel) bereits umgestiegen oder befinden sich im Prozesses des Umstiegs von PBA auf CBA. Findet innerhalb dieser Studien ein Moduswechsel statt, dann muss die Vergleichbarkeit zwischen den Ergebnissen der unterschiedlichen Administrationsmodi gewährleistet werden. Unterschiede in den Eigenschaften der Modi, wie beispielsweise im Antwortformat, können sich dabei auf die psychometrischen Eigenschaften der Tests auswirken und zu sogenannten Modus-Effekten führen. Diese Effekte wiederum können sich in Unterschieden zwischen den Testscores widerspiegeln, sodass diese nicht mehr direkt miteinander vergleichbar sind. Die zentrale Frage dabei ist, ob es durch den Moduswechsel zu einer Veränderung des gemessenen Konstruktes kommt. Ist dies der Fall, so können Testergebnisse aus unterschiedlichen Administrationsmodi nicht miteinander verglichen und die Ergebnisse aus dem computerbasierten Test nicht analog zu den Ergebnissen aus dem papierbasierten Test interpretiert werden. Auch Veränderungen, die aus Messungen zu verschiedenen Zeitpunkten und mit unterschiedlichen Modi resultieren, lassen sich dann nicht mehr beschreiben. Es kann jedoch auch Modus-Effekte geben, die zwar nicht das gemessene Konstrukt betreffen, aber sich beispielsweise in der Schwierigkeit der Items niederschlagen. Solange aber das erfasste Konstrukt bei einem Moduswechsel unverändert bleibt, können diese Modus-Effekte bei der Berechnung der Testscores berücksichtigt und die Vergleichbarkeit gewährleistet werden. Somit ist, nicht nur im Hinblick auf gültige Trendschätzungen, der Analyse von Modus-Effekten ein hoher Stellenwert beizumessen.

Da die bisherige Befundlage in der Literatur zu Modus-Effekten sowohl hinsichtlich der Stärke der gefundenen Effekte, als auch in Bezug auf die verwendeten Methoden sehr heterogen ist, ist das Ziel des ersten Beitrags dieser publikationsbasierten Dissertation, eine Anleitung für eine systematische Durchführung einer Äquivalenzuntersuchung, speziell für

Large-Scale Assessments, zu geben. Dabei wird die exemplarisch dargelegte Modus-Effekt-Analyse anhand von zuvor definierten und in ihrer Bedeutsamkeit belegten Kriterien auf der Test- und Item-Ebene illustriert. Zudem wird die Möglichkeit beschrieben, auftretende Effekte anhand von Eigenschaften des Administrationsmodus', beispielsweise des Antwortformats oder der Navigationsmöglichkeiten innerhalb des Tests, zu erklären. Im zweiten und dritten Beitrag findet sich jeweils eine empirische Anwendung der im ersten Beitrag beschriebenen schematischen Modus-Effekt-Analyse mit unterschiedlicher Schwerpunktsetzung. Dazu wurden die Daten eines Leseverständnistests aus der Nationalen Begleitforschung von PISA 2012 sowie zweier Leseverständnistests im NEPS, die jeweils sowohl papier- als auch computerbasiert administriert wurden, analysiert. Das Kriterium der Konstrukt-Äquivalenz steht dabei als wichtigstes Äquivalenz-Kriterium im Fokus. Zusätzlich wurde Äquivalenz in Bezug auf die Reliabilität und die Item-Parameter (Schwierigkeit und Diskrimination) untersucht. Im zweiten Beitrag wurden darüber hinaus interindividuelle Unterschiede im Modus-Effekt in Bezug zu basalen Computerfähigkeiten und zum Geschlecht gesetzt. Der dritte Beitrag fokussiert die Item-Eigenschaften, die als mögliche Quellen von Modus-Effekten herangezogen werden können und bezieht diese zur Erklärung von Modusunterschieden in die Analyse mit ein. In beiden Studien wurde keine Evidenz gefunden, dass sich das Konstrukt bei einem Wechsel des Administrationsmodus ändert. Lediglich einzelne Items wiesen am Computer im Vergleich zum PBA eine erhöhte Schwierigkeit auf, wobei sich der größte Teil der Items als invariant zwischen den Modi erwies. Für zwei Item-Eigenschaften wurde ein Effekt auf die erhöhte Schwierigkeit der Items am Computer gefunden. Interindividuelle Unterschiede im Modus-Effekt konnten nicht durch basale Computerfähigkeiten oder das Geschlecht erklärt werden.

Diese Dissertation leistet einen wesentlichen Beitrag zur Systematisierung von Äquivalenzuntersuchungen, insbesondere solchen in Large-Scale Assessments, indem sie die wesentlichen Kriterien für die Beurteilung von Äquivalenz herausstellt und diskutiert sowie deren Analyse methodisch aufbereitet. Die Relevanz von Modus-Effekt Studien wird dabei nicht zuletzt durch die Ergebnisse der beiden empirischen Beiträge hervorgehoben. Schließlich wird der Bedeutung des Einbezugs von Item-Eigenschaften hinsichtlich der Beurteilung der Äquivalenz Ausdruck verliehen.

Auflistung der Einzelbeiträge

Beitrag I: Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The Transition to Computer-Based Testing in Large-Scale Assessments: Investigating (Partial) Measurement Invariance between Modes. *Psychological Test and Assessment Modeling, 58* (4), 587-606.

Beitrag II: Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (*submitted, Computers in Human Behavior*). Construct Equivalence of PISA Reading Comprehension Measured with Paper-based and Computer-based Assessment.

Beitrag III: Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (*submitted, Educational and Psychological Measurement*). What makes the difference? The Impact of Item Properties on Mode Effects in Reading Assessments.

1. Einleitung

In der Bildungsforschung und ihren nationalen wie internationalen Studien war Papier lange Zeit das Medium für Assessments. Daten wurden durch das Beantworten, Ankreuzen und Markieren von Fragen erhoben, die auf Papier gedruckt und zu Testheften zusammengebunden waren. Mit dem Erreichen des digitalen Zeitalters vollzieht sich der technologische Wandel auch bei der Anwendung von Assessments. Der Computer ist „zu einem unverzichtbaren Werkzeug pädagogischer und psychologischer Diagnostik geworden" (Hartig, Kröhne, & Jurecka, 2007, S. 57). Damit hält er auch Einzug in die großen Bildungsvergleichsstudien wie PISA (Programme for International Student Assessment; OECD, 2016) und nationalen Studien über Bildungsverläufe und -entwicklungen wie NEPS (Nationales Bildungspanel; Artelt, Weinert, & Carstensen, 2013). Findet innerhalb dieser Studien ein Umstieg von papierbasiertem (PBA) auf computerbasiertes Assessment (CBA) statt, so muss die Vergleichbarkeit zwischen den Administrationsmodi gewährleistet werden (vgl. Sälzer & Reiss, 2016). Unterschiede in den Eigenschaften der Modi können dabei die psychometrischen Eigenschaften der Tests bzw. der Items[1] beeinflussen und zu sogenannten Modus-Effekten führen (vgl. Kröhne & Martens, 2011). Diese Effekte wiederum können sich in Unterschieden zwischen den Testscores widerspiegeln, sodass diese nicht direkt miteinander vergleichbar sind. Die zentrale Frage dabei ist, ob es durch den Wechsel auf CBA zu einer Veränderung des gemessenen Konstruktes im Sinne einer Zunahme von konstrukt-irrelevanter Varianz kommt (vgl. Huff & Sireci, 2001). Ist dies der Fall, so lassen sich Testergebnisse aus unterschiedlichen Administrationsmodi nicht miteinander vergleichen und analog interpretieren. Auch Veränderungen, resultierend aus Messungen zu verschiedenen Zeitpunkten und mit unterschiedlichen Modi, können dann nicht mehr als solche interpretiert werden. Es kann dabei aber auch solche Modus-Effekte geben, die zwar nicht das gemessene Konstrukt betreffen, sich aber beispielsweise in der Schwierigkeit der Items niederschlagen. Solange jedoch das erfasste Konstrukt bei einem Moduswechsel unverändert bleibt, können diese Modus-Effekte bei den Berechnungen der Testscores berücksichtigt und Vergleichbarkeit gewährleistet werden (vgl. Heine et al., 2016). Somit ist, nicht nur im Hinblick auf gültige Trendschätzungen, der Analyse von Modus-Effekten ein hoher

[1] Unter Items sollen hier gemäß der Definition von Rost (2004) „die Bestandteile eines Tests, die eine Reaktion oder Antwort hervorrufen sollen, also die Fragen, Aufgaben" verstanden werden (S. 18).

Stellenwert beizumessen (Robitzsch et al., 2016). Dabei wirft die Notwendigkeit der Gewährleistung gültiger Vergleiche zwischen PBA und CBA relevante Forschungsfragen auf, die sich auf die Äquivalenz zwischen den Testformen beziehen.

Die vorliegende publikationsbasierte Dissertation befasst sich mit Fragen der Vergleichbarkeit von PBA und CBA sowie mit den dafür relevanten Analyseschritten. Im nächsten Kapitel wird dazu zunächst ein Überblick über die Vorteile und Herausforderungen des computerbasierten Testens gegeben und die Relevanz von Modus-Effekt-Studien herausgestellt. Im Anschluss wird der Forschungsgegenstand der Dissertation hergeleitet und die drei Einzelbeiträge, die in der Vollversion im Anhang zu finden sind, in einer Zusammenfassung dargestellt. Der Rahmentext der Dissertation wird durch eine kritische Reflexion und Diskussion sowie einen Ausblick abgeschlossen.

2. Computer in Assessments

2.1 Vorteile und Herausforderungen von computerbasiertem Assessment

Die Vorteile des computerbasierten Assessments sind vielfältig. Die großen Vorteile aus ökonomischer Sicht sind die entfallenden Kosten für den Papierdruck, die Lagerung des Datenmaterials sowie der Wegfall der händischen Auswertung (Pomplun, Frey, & Becker, 2002). Aber auch aus diagnostischer Sicht gibt es wesentliche Aspekte, die für computerbasiertes Testen sprechen. So sorgt die softwarebasierte und automatisierte Auswertung für ein schnelleres oder sogar unmittelbares Feedback (Bennett, 2003; Bodmann & Robinson, 2004; Pomplun et al., 2002), menschliche Fehlerquellen bei der Auswertung werden reduziert und die Standardisierung der Auswertung wird erhöht (Jurecka & Hartig, 2007; Wang, Jiao, Young, Brooks, & Olson, 2008). Computerbasiertes Testen ermöglicht darüber hinaus ein höheres Maß an Messeffizienz, denn die Möglichkeiten der Nutzung der neuen Technologie gehen über die Anwendungen heraus, die beim papierbasierten Testen Standard waren (Bennett, Jenkins, Persky, & Weiss, 2003; Davey, 2011), oder schaffen sogar gänzlich neue Anwendungsfelder (z.B. Information and Communication Technology Literacy, kurz ICT-Literacy; Frey & Hartig, 2013). Neue und innovative Aufgabenformate sowie die interaktive Gestaltung von Tests tragen zur Verbesserung der Präsentation und der validen Erfassung des jeweils zu messenden Konstruktes bei (Jurecka & Hartig, 2007;

Parshall, Harmes, Davey, & Pashley, 2010; Sireci & Zenisky, 2006). Durch programmierbare Algorithmen werden Multi-Stage und adaptives Testen einfacher und beispielsweise auch in Gruppentest-Settings einsetzbar (vgl. Kröhne & Martens, 2011). Die Verfügbarkeit von Log- und Prozess-Daten, die man durch die Sammlung von Hintergrunddaten beschreiben kann, stärken das analytische Potenzial dadurch, dass zum Beispiel genaue Zeitdaten (Reaktions- und Bearbeitungszeiten) zusätzlich erfasst und ebenfalls ausgewertet werden können (Goldhammer, Naumann, Rölke, Stelter, & Tóth, 2017; Noyes & Garland, 2008). Auch bestimmte Interaktionsmuster im Bearbeitungsprozess können anhand dieser miterhobenen Daten identifiziert werden, was sich in papierbasierten Erhebungen nur mit viel Aufwand realisieren lässt (Frey & Hartig, 2013). Diese zusätzlichen Informationen helfen dabei Testbearbeitungsprozesse besser zu verstehen und die Qualität der erhobenen Daten zu überprüfen.

Auch motivationale Faktoren auf Seiten der Testteilnehmer[2] sprechen für computerbasiertes Testen. In verschiedenen Studien zeigte sich eine Präferenz für die Teilnahme an Tests am Computer (Choi & Tinkler, 2002; Higgins, Russel, & Hoffmann, 2005; Noyes & Garland, 2008; Wang, 2004) und gleichzeitig eine erhöhte Motivation für die Testteilnahme (Nikou & Economides, 2016).

Bei all den Vorteilen, die der Einsatz der Computertechnologie in Assessments mit sich bringt, entstehen aber auch neue Herausforderungen, die es zu bewältigen gilt (vgl. Hartig et al., 2007; Noyes & Garland, 2008). Dazu gehört, dass eine neue Infrastruktur geschaffen und zur Verfügung gestellt werden muss, um computerbasiertes Testen zu ermöglichen. Die Testerstellung am Computer erfordert entsprechende Werkzeuge, wie die Bereitstellung geeigneter Programme zur Computerisierung sowie eine entsprechende virtuelle Umgebung zur Speicherung und zum Abruf eines Tests. Unter Einsatz der Programme zur Umsetzung computerbasierter Tests lassen sich dann einige Testaufgaben leichter und andere eher schwieriger computerisieren. Dabei kann die Umsetzung von einfachen Multiple-Choice-Aufgaben einfacher realisiert werden als die von komplexeren, beispielsweise geometrischen Zeichenaufgaben. Aber nicht nur im Hinblick auf die Gestaltung, sondern auch auf den Einsatz und die Durchführung des Tests müssen entsprechende technische Voraussetzungen erfüllt sein. So müssen Zugriffsrechte berücksichtigt und gleichzeitig der Datenschutz sowohl personen- als auch testseitig gewährleistet werden (Hartig, et al., 2007). Auch die Verfügbarkeit von Hardware, auf

[2] Zum Zweck der besseren Lesbarkeit wird im Fließtext bewusst auf die zusätzliche Anführung der weiblichen Form verzichtet; sämtliche Angaben beziehen sich natürlich auf beide Geschlechter.

welcher der Test bearbeitet werden soll (vgl. Heine et al., 2016) sowie der Zugang zum Internet im Fall von Online-Testungen (vgl. Cronk & West, 2002) stellen Herausforderungen dar, denen sich angenommen werden muss. Diese sind, vor allem im Gruppentest-Setting an Schulen, nicht immer einfach und kostenminimierend zu bewältigen (Jurecka & Hartig, 2007).

2.2 Effekte des Administrationsmodus

Soll ein bestehender Papier-Test auf den Computer übertragen und die Ergebnisse aus PBA und CBA verglichen werden, so besteht die Herausforderung vor allem darin, die computergestützte Umsetzung möglichst nah anzulehnen an den Papier-Test. Dadurch können Modus-Effekte zwar verringert, jedoch keinesfalls gänzlich ausgeschlossen werden, denn es gibt eine Reihe von Eigenschaften, in denen sich die beiden Modi unterscheiden (vgl. Kröhne & Martens). Diese Unterschiede lassen sich im Wesentlichen auf das veränderte Testformat (Testheft in DIN A 4-Format entgegen einem Computerbildschirm) sowie die unterschiedliche Handhabung (Bearbeitung mit dem Stift im Gegensatz zur Computer-Maus oder zum Touchpad) zurückführen. Gibt es solche veränderte Eigenschaften der Testadministration, dann können sich diese auf psychometrische Eigenschaften des Tests, wie das gemessene Konstrukt im Sinne des Anteils konstrukt-irrelevanter Varianz und die Reliabilität, aber auch auf die Parameter der Items (z.B. Diskrimination und Schwierigkeit) auswirken und damit Modus-Effekte bedingen (vgl. Mead & Drasgow, 1993; Puhan, Boughton, & Kim, 2007). Dabei steigt deren Wahrscheinlichkeit mit der Komplexität der Darstellung und Bedienbarkeit des Tests am Computer (Pommerich, 2004). Modus-Effekte können also durch ein Konglomerat von unterschiedlichen Eigenschaften der Testformen begründet sein und sich auf die Testergebnisse und nicht zuletzt auf deren Vergleichbarkeit auswirken. Sind Unterschiede zwischen Testergebnissen das Resultat von Modus-Effekten, so kann nicht zwingend gewährleistet werden, dass unabhängig davon, in welchem Modus der Test bearbeitet wurde, gleich fähige Personen auch den gleichen Testscore erhalten (Raju, Laffitte, & Byrne, 2002; Van den Noortgate & De Boeck, 2005).

Mit Blick auf die Testfairness müssen darüber hinaus differentielle Effekte des Administrationsmodus‘ berücksichtigt werden. Unterschiede in der Fähigkeit oder der Vertrautheit im Umgang mit dem Computer dürfen sich dabei nicht nachteilig auf die Ergebnisse der Testpersonen auswirken (vgl. Jurecka & Hartig, 2007). Entsprechende Maße sollten daher in den jeweiligen Studien miterfasst werden, um interindividuelle Unterschiede

im Modus-Effekt in Abhängigkeit relevanter personenbezogener Eigenschaften beurteilen zu können.

Das Ziel einer Modus-Effekt-Studie ist es, die Äquivalenz zweier Testformen (hier zwischen einem papier- und einem computerbasierten Test) anhand klar definierter Äquivalenzkriterien zu beurteilen (siehe auch den Beitrag I der vorliegenden Dissertation). Der intendierte Vergleich sowie die gewünschte Interpretation der Testscores bestimmen dabei diejenigen Kriterien, die zwischen den Testformen äquivalent sein sollen. Insofern kann eine Modus-Effekt-Studie auch als Teil der Validierung im Sinne der Testscore-Interpretation verstanden werden (AERA, APA, & NCME, 2014). Als wichtigstes Äquivalenzkriterium im Hinblick auf die Interpretation und die Vergleichbarkeit von Testscores sowohl auf der Populations- als auch der Individualebene ist die Konstrukt-Äquivalenz zu nennen. Mit diesem Kriterium wird geprüft, ob durch den Wechsel zu CBA das gleiche Konstrukt wie mit dem papierbasierten Test erhoben wird oder ob es zu einer Zunahme konstrukt-irrelevanter Varianz kommt (AERA, APA, & NCME, 2014; Huff & Sireci, 2001). Ist Letzteres der Fall, so kann die Interpretation der Testscores des papierbasierten Tests nicht auf die des Computer-Tests übertragen und somit die Testergebnisse nicht verglichen werden (siehe auch den Beitrag I der vorliegenden Dissertation). Als weiteres Kriterium zur Beurteilung der Äquivalenz zweier Tests sollte die Reliabilität herangezogen werden (AERA, APA, & NCME, 2014; Holland & Dorans, 2006; ITC, 2005). Bei Large-Scale-Assessments (LSA) ist dieses Kriterium vor allem in Längsschnittstudien bedeutsam, also dann, wenn ein Linking zwischen verschiedenen Zeitpunkten und Administrationsmodi erforderlich ist (Holland & Dorans, 2006; siehe auch den Beitrag III der vorliegenden Dissertation). Als weitere Äquivalenzkriterien können die Item-Diskrimination und Item-Schwierigkeit herangezogen werden. Modus-Effekte, die sich als Unterschiede zwischen den Item-Parametern (in getrennter Berechnung für Diskriminationen und Schwierigkeiten) zeigen, lassen sich klassifizieren in a) homogene Effekte, die sich gleichermaßen auf alle Items auswirken, oder b) heterogene Effekte, die sich nur für einzelne Items zeigen. Dabei kann weiter spezifiziert werden, ob sich diese heterogenen Effekte systematisieren, also durch bestimmte Item-Eigenschaften (z.B. das Antwortformat) oder sonstige spezifische Unterschiede in der Umsetzung der Testversionen, erklären lassen (vgl. Green, Bock, Humphreys, Linn, & Reckase, 1984).

Mit der Analyse und dem Wissen um die Existenz von Modus-Effekten kann diesen Rechnung getragen werden. Besteht eine ausreichende Evidenz, dass sich das gemessene

Konstrukt durch einen Moduswechsel nicht ändert, können Item-Parameter (z.B. Diskrimination und Schwierigkeit) dementsprechend adjustiert und die Testscores unter Berücksichtigung der aufgetretenen Effekte berechnet werden (vgl. Heine et al., 2016). Sind beispielsweise einzelne CBA-Items von einem Modus-Effekt betroffen, so muss für jedes betroffene CBA-Item ein modusspezifischer Shift-Parameter berechnet werden, welcher die Verschiebung gegenüber dem PBA-Item-Parameter berücksichtigt. Im Fall einer homogenen Verschiebung aller CBA-Item-Parameter kann ein modusspezifischer Gesamt-Shift-Parameter berechnet und auf alle Items angewendet werden (siehe auch den Beitrag I der vorliegenden Dissertation). Durch eine solche Adjustierung der von Modus-Effekten betroffenen Items ist es möglich, eine gemeinsame Metrik zu bilden und die Testscores aus den unterschiedlichen Modi trotz der vorhandenen Modus-Effekte zu vergleichen. Somit kommt der Analyse von Modus-Effekten bei einem Wechsel des Administrationsmodus und hinsichtlich der Gewährleistung von Vergleichbarkeit zwischen den Modi höchste Bedeutung zu (AERA, APA, & NCME, 2014; ATP, 2000; ITC, 2005).

2.3 Modus-Effekt-Studien – Relevanz bis heute

Die ersten Modus-Effekt-Studien wurden in den achtziger Jahren des zwanzigsten Jahrhunderts durchgeführt (z. B. Gould & Grischkowsky, 1984; Mazzeo & Harvey, 1988). Unterschiede wurden damals vor allem auf die niedrige Auflösung des Computerbildschirms und die damit verbundene schlechtere Lesbarkeit und schnellere Ermüdung der Augen zurückgeführt. Diese Faktoren können aufgrund der technischen Weiterentwicklung und des heutigen Stands der Computertechnologie weitestgehend ausgeschlossen werden. Jedoch wirft die durch die neuen Technologien ermöglichte und immer breiter werdende Gestaltungsvielfalt eines Tests am Computer neue Fragen nach den Unterschieden zum Papier-Test auf, denen sich die Äquivalenzforschung widmen sollte. Sichtet man dazu die Literatur zu den Äquivalenzstudien der letzten 30 Jahre, dann zeigt sich eine sehr heterogene Befundlage. Einige Forscher konnten Äquivalenz aufzeigen, andere dagegen fanden erhebliche Unterschiede zwischen den psychometrischen Eigenschaften der Testformen sowie den daraus resultierenden Testergebnissen (vgl. Noyes & Garland, 2008; Wang et al., 2008). Im Kontext von LSA zeigen auch die aktuellen Befunde aus dem im Jahr 2014 durchgeführten PISA Feldtest, dass Modus-Effekte auf internationaler (ETS, 2015) sowie nationaler Ebene (für Deutschland; siehe Robitzsch et al., 2016) existieren und die Trendschätzungen aus PISA 2015 infolgedessen unter Berücksichtigung der Effekte des Moduswechsels interpretiert werden müssen. Nur selten beziehen Forschungsarbeiten in ihre

Äquivalenzanalysen auch konkret die Test- und Item-Eigenschaften ein, die für Modus-Effekte verantwortlich sein können bzw. sich zwischen den Modi unterscheiden (Pommerich, 2004). Bisher konnte vor allem gezeigt werden, dass das Scrollen in Lesetexten und Aufgaben am Computer deren Schwierigkeit erhöht (z. B. Bridgeman, Lennon, & Jackenthal, 2001; Higgins et al., 2005; Kim & Huynh, 2008; Pommerich, 2004). Jedoch gibt es auch Befunde, die nahelegen, dass das Scrollen keinen signifikanten Einfluss auf die Schwierigkeit hat (Yamamoto, 2012). Das Antwortformat und seine Komplexität scheinen ebenfalls potenzielle Quellen von Modus-Effekten zu sein (Parshall, Spray, Kalohn, & Davey, 2002). Dabei konnte zwischen papierbasierten und computerbasierten Aufgaben mit Multiple-Choice Antwortformat eher Äquivalenz nachgewiesen werden (Bennett et al., 2008; Bodmann & Robinson, 2004; Parshall et al., 2002), als dies bei komplexeren Aufgaben, wie beispielsweise bei Zuordnungsaufgaben, der Fall war. Letztere erwiesen sich im Vergleich zur herkömmlichen Bearbeitung im Testheft als schwieriger, wenn sie am Computer mittels Drop-Down-Boxen dargestellt wurden (Heerwegh & Loosveldt, 2002; siehe auch den Beitrag III der vorliegenden Dissertation). Bei der Betrachtung von personenbezogenen Eigenschaften im Zusammenhang mit interindividuellen Unterschieden im Modus-Effekt sind bisherige Ergebnisse ebenfalls nicht konsistent (vgl. Clariana & Wallace, 2002; Leeson, 2006). Eine systematische Benachteiligung von Personen mit geringerer Computerfähigkeit, weniger Erfahrung im Umgang mit Computern oder aufgrund ihres Geschlechts muss somit weder zwingend angenommen, noch kann sie grundsätzlich ausgeschlossen werden.

Eine große Heterogenität zeigt sich aber nicht nur in den Ergebnissen, sondern auch hinsichtlich der verwendeten Methoden. Dabei bestehen auch hinsichtlich der Stichprobenzahlen und damit bezogen auf die Power der statistischen Tests erhebliche Unterschiede zwischen den Studien. Während einige Forschungsarbeiten Äquivalenz lediglich auf der manifesten Ebene untersuchten, beispielsweise mithilfe von Mittelwertvergleichen oder Korrelationen von Item-Parametern (vgl. z. B. Alexander, Bartlett, Truell, & Ouwenga, 2001; Pomplun & Custer, 2005; Puhan, et al., 2007), drangen andere viel tiefer in die Struktur von Modus-Effekten vor, indem sie spezielle Item-Eigenschaften in die Analyse einbezogen (Bennett et al., 2008; Parshall et al., 2010). Häufig wurde auch die Äquivalenz des zu messenden Konstruktes vernachlässigt (Kim & Huynh, 2008), ohne deren Einbeziehung die Interpretation der Ergebnisse solcher Analysen nicht möglich ist. In den meisten Studien fehlte es darüber hinaus an klar definierten Kriterien, anhand derer Äquivalenz beurteilt wurde. Die fehlende Systematik der für die Untersuchung der Äquivalenz eingesetzten Methoden erschwert den direkten Vergleich zwischen den unterschiedlichen Studien und

ihren Ergebnissen. Sie könnte darüber hinaus für die uneinheitlichen Befunde zu Modus-Effekten verantwortlich sein (Schröders & Wilhelm, 2011; Wang et al., 2008).

Die Ergebnisse vorliegender Äquivalenzstudien können also in ihrer Grundgesamtheit bislang nicht generalisiert werden, und die Existenz von Modus-Effekten im Einzelfall nicht ausgeschlossen werden. Infolgedessen müssen die Effekte des Administrationsmodus in jedem Einzelfall überprüft werden, wenn die Testscores aus Tests in unterschiedlichen Modi verglichen werden sollen (AERA, APA, & NCME, 2014; Kröhne & Martens, 2011; Noyes & Garland, 2008; Wang et al., 2008). Die technischen Möglichkeiten des Einsatzes neuer Hardware zur Testadministration (z. B. Tablets und Smartphones; vgl. Illingworth, Morelli, Scott, & Boyd, 2015; Ling, 2016) erfordern darüber hinaus, die Untersuchung von Modus-Effekten auf andere Administrationsmodi auszuweiten. Dies unterstreicht die Bedeutsamkeit der kontinuierlichen Erforschung von Modus-Effekten auch in der Zukunft (vgl. Noyes & Garland, 2008). Auch die sich in diesem Zusammenhang vollziehenden Veränderungen der Test-Settings sollten einen Schwerpunkt zukünftiger Äquivalenzforschung bilden. Das sowohl für papier- als auch für computerbasierte Testungen häufig gewählte Gruppentest-Setting wird ergänzt um die Möglichkeiten von Online-Testungen, deren Vergleichbarkeit ebenso gesichert werden muss (Ihme et al., 2009).

3. Der Forschungsgegenstand der Dissertation

Ausgehend von der großen Heterogenität, die sich in den Befunden und den Methoden von Modus-Effekten-Studien zeigt, soll die vorliegende Dissertation einen Beitrag dazu leisten, dieses Forschungsfeld zukünftig systematischer zu gestalten. Im Fokus stehen dabei die Kriterien, anhand derer Äquivalenz beurteilt werden soll. Die Definition dieser Kriterien legt die Grundlage für eine Modus-Effekt-Studie und dient dazu, die gewünschten Vergleiche und Interpretationen der Testscores vorzubereiten. Darüber hinaus tragen sie aber auch zur Transparenz und Systematik von Modus-Effekt-Studien bei. Da sich in den bisherigen Studien auch methodisch eine große Heterogenität zeigt, zielt die vorliegende Arbeit außerdem darauf ab, zu klären, welche Methoden und Untersuchungsdesigns für die Analyse der Effekte des Administrationsmodus geeignet sind. Dabei wird unter anderem aufgezeigt, in welchem Zusammenhang verschiedene Aspekte des Studiendesigns mit den Methoden stehen, mittels derer sich Äquivalenz untersuchen lässt, und welche Aussagen auf dieser Grundlage

möglich sind. Auch dient dies dem Ziel, die zukünftige Forschung zum Thema Modus-Effekte zu vereinheitlichen und vergleichbarer zu machen und dadurch einen Beitrag in Richtung der Generalisierbarkeit der Ergebnisse zu leisten.

Eine zentrale Rolle in der vorliegenden Arbeit spielt die Prüfung des Kriteriums der Konstrukt-Äquivalenz. Denn nur dann, wenn das gemessene Konstrukt durch den Moduswechsel nicht verändert wird, kann eine Vergleichbarkeit zwischen PBA und CBA auch retrospektiv (z. B. unter Adjustierung der Item-Parameter) hergestellt werden. Dieses Kriterium wird empirisch anhand von drei Leistungstests zum Leseverständnis geprüft. Zwei weitere Äquivalenzkriterien, die in dieser Arbeit ebenfalls empirisch daraufhin überprüft werden, ob sich Unterschiede zwischen den Administrationsmodi zeigen, sind die Reliabilität und die Item-Parameter (z. B. Diskrimination und Schwierigkeit der Items). Darüber hinaus werden personenbezogene Eigenschaften zur Erklärung von interindividuellen Unterschieden im Modus-Effekt in die Analyse einbezogen.

Neben der Analyse der Unterschiede in den Äquivalenzkriterien der beiden Testformen ist es auch notwendig, diejenigen Eigenschaften der Tests und Items zu identifizieren, welche die Wahrscheinlichkeit von Modus-Effekten erhöhen (Leeson, 2006). Die vorliegende Arbeit stellt dazu spezifische Item-Eigenschaften unter Einbezug empirischer Befunde als mögliche Quellen von Modus-Effekten heraus. Die Identifikation solcher Eigenschaften verhilft dabei nicht nur zu einem tieferen Verständnis der Effekte des Computers als Administrationsmodus, sondern dient auch der stetigen Verbesserung und der Anreicherung von Wissen über die Computerisierung von Papier-Tests.

4. Darstellung der Einzelbeiträge

Im Folgenden werden die drei Einzelbeiträge zusammenfassend dargestellt. Der Schwerpunkt liegt dabei auf den Zielen des jeweiligen Beitrages und, im Fall des zweiten und dritten Beitrags auf den empirischen Ergebnissen.

4.1 Beitrag I: Der Übergang von papierbasiertem zu computerbasiertem Assessment

Der erste Beitrag[3] reagiert auf die oben formulierte Kritik an der fehlenden Systematik vorliegender Modus-Effekt-Studien. Dabei besteht das Ziel darin, eine Anleitung für die Durchführung einer Äquivalenzuntersuchung (hier speziell für LSA) zu geben, die sich an bestimmten Kriterien orientiert und geeignete statistische Methoden anwendet. Dieser theoretisch-methodische Beitrag illustriert eine Modus-Effekt-Analyse anhand von vorher definierten und in ihrer Bedeutsamkeit belegten Kriterien auf der Test- und Item-Ebene und beschreibt darüber hinaus die Möglichkeit, auftretende Effekte anhand von Eigenschaften des Administrationsmodus‘ zu erklären.

4.1.1 Zusammenfassung

Design von Modus-Effekt-Studien. Bei der Planung einer Modus-Effekt-Studie kommt dem Studiendesign eine besondere Bedeutung zu, vor allem hinsichtlich der methodischen Überlegungen. Dabei ist zunächst die Entscheidung zu treffen, ob der Modus innerhalb einer Person (Inner-Subjekt-Design) oder zwischen Personen (Zwischen-Subjekt-Design) variiert werden soll. Hier ist, aufgrund der höheren statistischen Power, das Erstere vorzuziehen, nämlich dass alle Personen den Test in beiden Modi bzw. je einen Teil des Tests in je einem Modus bearbeiten (Schröders & Wilhelm, 2011). Konstrukt-Äquivalenz kann dann direkt mit Hilfe sogenannter Cross-Mode-Korrelationen (vgl. Mead & Drasgow, 1993) analysiert werden. In einem Zwischen-Subjekt-Design ist es hingegen möglich, Konstrukt-Äquivalenz aufgrund der Beziehung zu anderen Variablen (z. B. ein anderer Test der gleichen Domäne; vgl. ITC, 2005) zu beurteilen. Ein weiterer wichtiger Aspekt ist die Reihenfolge der Testteile. In einem Inner-Subjekt-Design sollte diese möglichst ausbalanciert sein und die Teilnehmer dem Modus randomisiert zugeordnet werden (z.B. Pomplun et al., 2002). Liegt ein Zwischen-Subjekt-Design vor, so muss zumindest die Randomisierung der Testteilnehmer zu den Modi gewährleistet werden. Denn nur dann, wenn die Modus-Gruppen zufallsäquivalent sind, kann ausgeschlossen werden, dass die gefundenen Modus-Effekte durch Gruppenunterschiede in dem gemessenen Konstrukt oder anderen personenbezogenen Eigenschaften zustande kommen (Osterlind & Everson, 2009). Wird die Bearbeitungsreihenfolge von den Testteilnehmern selbst gewählt, so sollten solche Variablen miterfasst werden, welche Rückschlüsse auf den Entscheidungsprozess ermöglichen, damit

[3] Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The Transition to Computer-Based Testing in Large-Scale Assessments: Investigating (Partial) Measurement Invariance between Modes. *Psychological Test and Assessment Modeling, 58* (4), 587-606.

sich Zufallsäquivalenz retrospektiv erzeugen lässt (vgl. Hox, de Leeuw, & Zijlmans, 2015). An dieser Stelle ist auch eine Abgrenzung von Modus-Effekt-Studien zu derjenigen Forschung zu ziehen, die sich mit Differential Item Functioning (DIF) beschäftigt (vgl. Reise, Widaman, & Pugh, 1993). In DIF Studien soll Äquivalenz zwischen Gruppen untersucht werden, die sich in bedeutsamen personenbezogenen Merkmalen, beispielsweise dem Geschlecht oder der Muttersprache, unterscheiden. Diese Gruppen sind also per se nicht zufallsäquivalent und die Vergleichbarkeit muss auf anderem Wege (z. B. über Anker-Items) hergestellt werden. Modus-Effekt-Studien hingegen weisen durch die realisierte Möglichkeit der randomisierten Zuweisung der Teilnehmer zu Modus-Gruppen einen experimentellen Charakter auf, sodass die Gruppen, wenn sie erfolgreich randomisiert wurden, als zufallsäquivalent betrachtet werden können.

Äquivalenzkriterien. Neben der Festlegung des Studiendesigns gilt die Aufmerksamkeit den Überlegungen zur statistischen Auswertung der Studie bzw. der Beantwortung der Äquivalenz-Fragestellung. Die Modus-Effekt-Analyse sollte sich dabei an klar definierten Äquivalenzkriterien orientieren. Diese Kriterien wiederum leiten sich aus den intendierten Vergleichen der Testscores ab. Dabei verlangen Vergleiche zwischen den Modi sowohl auf der Individual- als auch auf der Populationsebene zumindest, dass sichergestellt wird, dass das Konstrukt, das erfasst und auf dessen Grundlage Aussagen getroffen werden sollen, sich nicht ändert, also keine konstrukt-irrelevante Varianz hinzukommt, wenn der Modus gewechselt wird (AERA, APA, & NCME, 2014; Huff & Sireci, 2001; ITC, 2005; Parshall et al., 2002; Penfield & Camilli, 2007; Puhan et al., 2007; Russell, Goldberg, & O'Connor, 2003). Dies führt zum ersten und wichtigsten Äquivalenzkriterium, nämlich dem der Konstrukt-Äquivalenz. In dem Beitrag wird zur beispielhaften Darstellung einer Modus-Effekt-Analyse von zufallsäquivalenten Gruppen in einem Zwischen-Subjekt-Design ausgegangen. Zur Überprüfung des ersten Äquivalenzkriteriums wird daher die Analyse der Beziehung zu externen Variablen illustriert, orientiert an den von AERA, APA und NCME herausgegebenen Standards zur Validität (AERA, APA, & NCME, 2014) sowie den Richtlinien zu computerbasiertem Testen der Internationalen Testkommission (ITC, 2005). Als externe Variablen können hier zum Beispiel Tests in derselben oder in einer anderen Domäne herangezogen werden, die jeweils mit dem Computer- und auch mit dem Papier-Test korreliert werden. Konstrukt-Äquivalenz erfordert, dass sich diese Korrelationen nicht signifikant voneinander unterscheiden. Als zweites wird die Reliabilität der beiden Testversionen als Kriterium für die Beurteilung von Äquivalenz beschrieben (AERA, APA, & NCME, 2014; Holland & Dorans, 2006; ITC, 2005; Kolen & Brennan, 2004), und in einem

weiteren Schritt werden die Item-Parameter aus PBA und CBA als Äquivalenzkriterien herangezogen. Das bei der Vorbereitung geschätzte Skalierungsmodell bestimmt dabei die Parameter, in denen die Modus-Effekte abgebildet werden können. Wird beispielsweise das 1-PL-Modell (Rasch, 1960) zur Skalierung eingesetzt, dann können Modus-Effekte in den Item-Schwierigkeiten, und bei Verwendung eines 2-PL-Modells (Birnbaum, 1968) zusätzlich noch in den Item-Diskriminationen abgebildet werden. An dieser Stelle wird ein Mehrgruppen-IRT-Modell schematisch illustriert und ein neuer Parameter als Modus-Effekt definiert, welcher den Unterschied der Item-Parameter (im Beitrag beispielhaft illustriert anhand der Item-Schwierigkeit) aus PBA und CBA abbildet. Zeigt sich auf der Ebene des Gesamttests in der Summe der Differenzen aller Item-Parameter ein Modus-Effekt, der signifikant von Null verschieden ist, so lässt sich schlussfolgern, dass einzelne Schwierigkeitsverschiebungen in den Items sich nicht gegenseitig kompensieren. Daran anschließend sind die Modus-Effekte auf der Item-Ebene zu analysieren und die Differenzen der Item-Schwierigkeiten für jedes Item-Paar aus PBA und CBA hinsichtlich ihrer Signifikanz zu beurteilen. Zudem ist zu analysieren, ob alle Items von einer Schwierigkeitsverschiebung betroffen sind oder nur einzelne. Auf der Item-Ebene kommt den Item-Eigenschaften überdies eine besondere Bedeutung zu. Dazu wird eine Erweiterung des Modells vorgeschlagen, die es ermöglicht, weitere Item-Eigenschaften einzubeziehen und ihren Effekt auf die Modusunterschiede zu schätzen.

Konsequenzen aus Modus-Effekten. Zeigen sich auf der Item-Ebene für einzelne Items Modus-Effekte, wohingegen sich für andere die Item-Parameter zwischen den Modi nicht unterscheiden, so kann von partieller Invarianz gesprochen werden. Dabei verhält es sich so, dass sich beim Vorliegen von zufallsäquivalenten Gruppen und Äquivalenz des erfassten Konstruktes Schwierigkeitsverschiebungen einzelner und sogar aller Items durch Adjustierung der Item-Parameter ausgleichen lassen. Im Fall von partieller Invarianz erhalten die CBA-Items mit Modus-Effekt einen modusspezifischen Item-Parameter. Dieser wird bei einem Modus-Effekt, von dem alle Items betroffen sind, zu einem Parameter vereinfacht, welcher die Verschiebung aller CBA-Items berücksichtigt. Dadurch ist es möglich, eine gemeinsame Metrik für den Vergleich von Testscores aus unterschiedlichen Modi herzustellen.

4.2 Beitrag II: Konstrukt-Äquivalenz zwischen PBA und CBA

Der zweite Beitrag[4] nimmt die Konstrukt-Äquivalenz eines Leseverständnistests in den Fokus, der als PBA und CBA im Rahmen der Nationalen Begleitforschung von PISA 2012 in Deutschland, administriert wurde. Darüber hinaus werden interindividuelle Unterschiede im Modus-Effekt anhand von Personenmerkmalen zu erklären versucht.

4.2.1 Zusammenfassung

Modus-Effekte in PISA. Während in der Hauptstudie von PISA 2012 noch papierbasierte Tests zum Einsatz kamen, wurde in PISA 2015 in allen Domänen auf CBA umgestellt. Gerade im Hinblick auf Aussagen über Veränderungen in den gemessenen Kompetenzen und in Bezug auf gültige Trendschätzungen ist dabei wesentlich, dass potenzielle Effekte des Administrationsmodus mitberücksichtigt werden. Zur Vorbereitung des Umstiegs auf CBA in der Hauptstudie von PISA 2015 bearbeiteten in der PISA 2015 Feldstudie (PISA 2015 Field Trial, OECD, 2016) die Test-Teilnehmer in einem randomisierten Zwischen-Subjekt-Design die Tests entweder papier- oder computerbasiert. Die darauf basierende Modus-Effekt-Analyse (OECD, 2016, Annex 6) sollte die Fragen zur Vergleichbarkeit von PBA und CBA sowie mit den vorherigen (papierbasierten) PISA-Zyklen hinsichtlich gültiger Trendschätzungen beantworten. Die für einzelne Items gefundenen Modus-Effekte wurden dabei für alle Länder als identisch angesehen und in der Hauptstudie entsprechend adjustiert, obwohl Robitzsch et al. (2016) zeigen konnten, dass länderspezifische Modus-Effekte die Trendschätzungen für Deutschland verzerrt haben könnten.

Die in diesem Beitrag beschriebenen Analysen untersuchen Modus-Effekte in einem papier- und computerbasierten Test zum Leseverständnis (die Hauptdomäne in PISA 2018) aus der Nationalen Begleitforschung von PISA 2012 in Deutschland.

Daten und Testinstrumente. Zwei Cluster eines Leseverständnistests (OECD, 2009) mit insgesamt 37 Items wurden computerisiert und in einem Zwischen-Subjekt-Design von N = 856 fünfzehnjährigen Schülern an deutschen Schulen bearbeitet. Die Stichprobe wurde dabei randomisiert aus den an der PISA 2012 Hauptstudie teilnehmenden Schulen gezogen. Als zusätzliche Inner-Subjekt-Komponente bearbeitete eine Teilmenge (N = 440) der

[4] Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (*submitted*, *Computers in Human Behavior*). Construct Equivalence of PISA Reading Comprehension Measured with Paper-based and Computer-based Assessment.

Stichprobe das eine Cluster als PBA und das andere als CBA. Die Reihenfolge der Modi wurde balanciert, und die Teilnehmer wurden den Testgruppen randomisiert zugewiesen. Zusätzlich bearbeiteten alle Teilnehmer einen Test zu basalen Computerfähigkeiten (Basic Computer Skills [BCS]; Goldhammer, Naumann, & Keßel, 2013). Darüber hinaus wurde im CBA ein Missing-Value-Indikator (vgl. Wang et al., 2005) eingeführt, der die Testteilnehmer nach dem Beenden einer Unit an noch nicht bearbeitete Items erinnerte. Die Testteilnehmer konnten dann entscheiden, die Bearbeitung der Unit dennoch zu beenden oder aber zurück zu den unbearbeiteten Items zu gehen.

Methode und Ergebnisse. Die in dieser Studie zu prüfenden Hypothesen bezogen sich auf Konstrukt-Äquivalenz, auf Modus-Effekte in Bezug auf Item-Schwierigkeiten sowie auf interindividuelle Unterschiede im Modus-Effekt und ihre Erklärbarkeit durch personenbezogene Eigenschaften. Darüber hinaus wurde der Unterschied zwischen den Modi hinsichtlich fehlender Werte untersucht. Die Modus-Effekt-Analyse wurde unter Orientierung an dem im ersten Beitrag beschriebenen Vorgehen unter Anwendung der Statistik-Software R (R Core Team, 2014) und Mplus (Muthén & Muthén, 1998-2015) durchgeführt. Als Skalierungsmodell wurde (wie auch bei PISA 2015) ein Generalized Partial Credit Modell (GPCM, Muraki, 1992) gewählt. Die Item-Diskriminationen wurden dabei für jedes Item frei geschätzt, jedoch zwischen den Modi als invariant angenommen. Aufgrund der zusätzlichen Inner-Subjekt-Komponente konnte Konstrukt-Äquivalenz mittels einer Cross-Mode-Korrelation untersucht werden. Dabei wurde angenommen, dass die Korrelation beider Cluster in einem Modus (PBA) nicht signifikant verschieden von der Korrelation der Cluster in unterschiedlichen Modi (PBA und CBA) ist, wenn Konstrukt-Äquivalenz vorliegt. Da kein Test-Teilnehmer beide Cluster in nur einem Modus bearbeitete, wurde als Referenz die Korrelation der mit beiden Clustern identischen papierbasierten Leseaufgaben aus PISA 2009 (OECD, 2010) herangezogen. Der Wald-Test (Wald, 1943) zeigte keinen signifikanten Unterschied zwischen der Korrelation der papierbasierten Aufgaben aus PISA 2009 und der Korrelation zwischen den Modi aus PISA 2012. Des Weiteren zeigte sich hinsichtlich der Schwierigkeit keine homogene Veränderung für alle Items, es wurden jedoch vier Items identifiziert, die eine signifikante Schwierigkeitsverschiebung zwischen den Modi aufwiesen. Um interindividuelle Unterschiede im Modus-Effekt zu identifizieren, wurde eine latente Differenzvariable als Modus-Effekt-Faktor in das Modell eingeführt. Entsprechend dem Ergebnis, dass der Moduswechsel zu keiner homogenen Schwierigkeitsverschiebung des gesamten Tests führte, zeigte sich der Mittelwert des Modus-Effekt-Faktors nicht signifikant von Null verschieden, jedoch korrelierte der Faktor negativ mit der Lesefähigkeit. Die

Varianz des Modus-Effekt-Faktors konnte dabei weder über basale Computerfähigkeiten noch über das Geschlecht erklärt werden. Des Weiteren zeigten sich hypothesenkonform in Bezug auf fehlende Werte ein signifikanter Unterschied zwischen der Anzahl nicht bearbeiteter Items (*omitted items*) und kein signifikanter Unterschied zwischen der Anzahl nicht erreichter Items (*not reached items*) zwischen PBA und CBA.

Die Ergebnisse dieser Studie zeigen, dass die Einführung von CBA in der Nationalen Begleitforschung von PISA 2012 zu keinen substantiellen Veränderungen führte. So wurde kein Hinweis darauf gefunden, dass sich das Konstrukt zwischen den Modi unterschied oder eine Veränderung der Schwierigkeit für den gesamten Test herbeigeführt wurde. In Bezug auf das Gütekriterium der Testfairness zeigte sich kein Zusammenhang von interindividuellen Unterschieden im Modus-Effekt mit der Computerfähigkeit oder dem Geschlecht einer Person. Die Korrelation des Modus-Effekt-Faktors mit der Lesefähigkeit impliziert jedoch, dass der Modus-Effekt für Personen mit geringerer Lesefähigkeit stärker ausgeprägt war. Die meisten Items zeigten sich invariant zwischen den Modi und können somit herangezogen werden um eine gemeinsame Metrik für den Vergleich von Testscores zwischen PBA und CBA zu bilden.

4.3 Beitrag III: Der Einfluss von Item-Eigenschaften auf Modusunterschiede

Gegenstand des dritten Beitrages[5] ist die Durchführung einer Modus-Effekt-Studie anhand der Daten zweier Leseverständnistests im NEPS (Artelt et al., 2013). Schwerpunkt dabei sind die Eigenschaften, die sich bei der Administration zwischen dem papier- und dem computerbasierten Test unterscheiden. Diese werden zur Erklärung von gefundenen Modus-Effekten in die Analyse einbezogen.

4.3.1 Zusammenfassung

Eigenschaften der Administrationsmodi. In einer theoretischen Abhandlung werden in diesem Beitrag Eigenschaften, die sich zwischen den Administrationsmodi unterscheiden können, anhand früherer Befunde zu Modus-Effekten dargestellt und in Bezug zur Computerisierung der aktuellen Studie gesetzt. Dabei wurden das Item-Layout, die Navigationsrestriktionen innerhalb des Tests sowie das Antwortformat als potenzielle Quellen von Modus-Effekten diskutiert (vgl. auch Parshall et al., 2002) und endsprechende

[5] Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (*submitted, Educational and Psychological Measurement*). What makes the difference? The Impact of Item Properties on Mode Effects in Reading Assessments.

Eigenschaften in den untersuchten NEPS Leseverständnistests identifiziert. Die Tests beinhalteten drei unterschiedliche Antwortformate, nämlich a) klassische Multiple-Choice Items, b) ein Format zur Beurteilung mehrerer Statements auf ihre Richtigkeit mit dichotomen Antworten richtig oder falsch (Complex-Multiple-Choice) und c) Zuordnungsaufgaben, die am Computer mit Combo-Boxen (Drop-Down-Menüs) umgesetzt wurden. Letztere unterschieden sich hinsichtlich ihrer Anforderungen an die Bedienung erheblich zwischen den Modi. Während im Testheft lediglich ein Buchstabe in das entsprechende Feld eingetragen werden musste, um beispielsweise eine Überschrift einer Textpassage zuzuordnen, musste dieser Buchstabe am Computer durch das Öffnen eines Drop-Down-Menüs ausgewählt werden. Als weitere Eigenschaften, die sich zwischen den Modi unterschieden, wurden zwei verschiedene Navigationsrestriktionen identifiziert: Zum einen für Items an erster und zweiter Position einer Unit, die im Testheft auf derselben Doppelseite wie der Lesetext abgebildet waren, wohingegen am Computer zwischen diesen Items und dem Text navigiert werden musste. Zum anderen musste am Computer in manchen Units der Lesetext auf zwei Bildschirmseiten aufgeteilt werden, was ebenfalls eine zusätzliche Navigation erforderte. Darüber hinaus wurde der Missing-Value-Indikator (vgl. Wang et al., 2005) eingeführt. Da dieser auf der Benutzung der technischen Möglichkeiten am Computer basiert, stellt er eine Eigenschaft dar, die gezielt verändert wurde, um den Effekt dieser neuen Technologie zu untersuchen.

Daten und Testinstrumente. Die durchgeführten Analysen basieren auf Daten einer im Jahr 2012 durchgeführten Studie im NEPS (Artelt et al., 2013), die der wissenschaftlichen Begleitung des Umstiegs von papier- auf computerbasiertes Assessment diente. Dazu wurden zwei Leseverständnistests[6] für die 7. Klassenstufe (Gehrer, Zimmermann, Artelt, & Weinert, 2013) computerisiert und auf ihre Äquivalenz zu den bestehenden Papier-Tests geprüft. Die Stichprobe betrug N = 1163 Schüler der 7. Jahrgangsstufe in verschiedenen deutschen Bundesländern. Die Tests wurden in einem Zwischen-Subjekt-Design bearbeitet, das heißt, jeder Schüler bearbeitete jeweils nur einen der beiden Tests entweder computer- oder papierbasiert. Die Schüler wurden dem jeweiligen Modus (PBA bzw. CBA) randomisiert zugewiesen. Zusätzlich zu dem im Fokus stehenden Leseverständnistest bearbeiteten alle Schüler einen papierbasierten Leseverständnistest für die 5. Klassenstufe (Gehrer et al., 2013) sowie einen Test zu basalen Computerfähigkeiten (BCS; Goldhammer et al., 2013) am Computer.

[6] Da drei Units zwischen den Tests identisch waren, wurden die Daten der beiden Tests für die Analysen zu einem Datensatz zusammengefasst (mit geplanten fehlenden Werten in denen sich unterscheidenden Units).

Methode und Ergebnisse. Die in der beschriebenen Studie geprüften Hypothesen bezogen sich auf die Äquivalenz sowohl des gemessenen Konstruktes als auch der Reliabilität und der Item-Parameter. Zusätzlich wurden fehlende Werte zwischen den Modi untersucht. Als Skalierungsmodell wurde das GPCM mit frei geschätzten Item-Diskriminationen innerhalb sowie auch zwischen den Modi bestimmt. Die Modus-Effekt-Analysen wurden somit sowohl hinsichtlich der Item-Diskriminationen als auch hinsichtlich der Item-Schwierigkeiten durchgeführt. Darüber hinaus wurden die vorab identifizierten Item-Eigenschaften in die Analyse einbezogen und ihr Effekt auf die gefundenen Modusunterschiede analysiert. Das Vorgehen orientierte sich dabei an den im ersten Beitrag dargestellten Methoden unter Einsatz von R (R Core Team, 2014) und Mplus (Muthén & Muthén, 1998-2015). Konstrukt-Äquivalenz wurde mittels der Beziehung zu externen Variablen (Korrelationen mit dem Leseverständnistest für die 5. Klassenstufe in PBA sowie mit BCS) bestimmt. Der Wald-Test zum Vergleich der Korrelation aus PBA und CBA mit den externen Variablen zeigte weder für den Zusammenhang mit dem Leseverständnistest der 5. Klassenstufe noch für den Zusammenhang mit BCS einen signifikanten Unterschied zwischen den Modi und lieferte somit die Evidenz dafür, dass sich das gemessene Konstrukt zwischen den Modi nicht unterschied. Die Reliabilität (EAP-Reliabilität ermittelt mit TAM; Kiefer, Robitzsch, & Wu, 2016) zeigte sich ebenfalls invariant zwischen PBA und CBA. Überdies unterschieden sich die Diskriminationen der Items nicht signifikant zwischen PBA und CBA und wurden daher für die nachfolgenden Analysen zwischen den Modi gleichgesetzt. Hinsichtlich der Item-Schwierigkeiten zeigte sich erwartungsgemäß ein signifikanter Modus-Effekt über die Summe aller Differenzen der Item-Paare aus PBA und CBA hinweg. Dieser Modus-Effekt war, ebenfalls hypothesenkonform, nicht homogen für alle Items, sondern variierte zwischen den Items. Dabei wiesen 22 der 42 Items einen Modus-Effekt auf. Bei den in die weitere Analyse einbezogenen Item-Eigenschaften konnte zum Teil ein Effekt auf die Verschiebung der Item-Schwierigkeiten zwischen den Modi nachgewiesen werden. Die Antwortformate Multiple-Choice und Complex-Multiple-Choice hatten erwartungsgemäß keinen Einfluss auf die Item-Schwierigkeiten zwischen PBA und CBA. Ebenfalls gemäß den Erwartungen, erhöhten sich durch die Verwendung von Combo-Boxen für Zuordnungsaufgaben am Computer sowie durch die zusätzliche Navigation für das erste und zweite Item einer Unit die Schwierigkeiten dieser Items am Computer. Dabei zeigte sich entgegen der Annahme kein Effekt der Aufteilung der Lesetexte auf zwei Bildschirmseiten. Die Analyse der fehlenden Werte wiederum ergab hypothesenkonforme Ergebnisse. Die Anzahl der nicht bearbeiteten Items (*omitted items*) fiel für den computerbasierten Test

signifikant geringer aus. Dagegen zeigte sich kein signifikanter Unterschied zwischen den Modi in der Anzahl der nicht erreichten Items (*not reached items*) am Ende des Tests.

Als wesentliches Ergebnis dieser Äquivalenzuntersuchung ist festzustellen, dass Evidenz dafür gefunden wurde, dass das gemessene Konstrukt durch den Wechsel des Administrationsmodus' nicht verändert wurde. Zudem konnte gezeigt werden, dass es bestimmte Item-Eigenschaften gab, die einen Einfluss auf die Schwierigkeitsverschiebung zwischen PBA und CBA hatten. Da es sich in dieser Studie um zufallsäquivalente Gruppen handelte, ergibt sich als Konsequenz, dass sich die Testscores beider Modi trotz der gefundenen Modus-Effekte vergleichen lassen. Alle Items mit Modus-Effekt bzw. Item-Gruppen, beispielsweise alle Items mit Combo-Box Antwortformat, erhalten dann einen modusspezifischen Item-Parameter, der die Schwierigkeitsverschiebung berücksichtigt.

5. Diskussion

Die Heterogenität der Ergebnisse bisher publizierter Modus-Effekt-Studien zeigt, dass Äquivalenzuntersuchungen in jedem Einzelfall notwendig sind (vgl. Pommerich, 2004; Wang et al., 2008). Während dies im Hinblick auf die Feststellung der Äquivalenz für jede einzelne Studie durchaus nützlich ist, tragen bereits publizierte Studien einzeln sowie in ihrer Gesamtheit (z. B. Wang et al., 2008) betrachtend bislang nur wenig zur Generalisierbarkeit der Ergebnisse bei. Statt für jede Übertragung eines papierbasierten Tests auf einen Computer eine eigene Äquivalenzuntersuchung zu planen, die beantragt, finanziert und letztlich auch durchgeführt und ausgewertet werden muss, müsste die Forschung sich vielmehr denjenigen Fragen widmen, die sich bei kritischer Betrachtung aus den zugänglichen Studien ableiten lassen. Dabei hat die wesentliche Frage danach, warum die Befundlage zu Äquivalenz-Fragestellungen so heterogen ist, bislang wenig Beachtung gefunden. In diesem Zusammenhang erscheint es wichtig, sich vor allem mit folgenden Fragen systematisch und kritisch auseinanderzusetzen: Wie wird Äquivalenz in den jeweiligen Studien verstanden und zu welchem Zweck (im Sinne intendierter Vergleiche) wird diese ermittelt? Hinsichtlich welcher äußeren Eigenschaften unterscheiden sich der papierbasierte und der daraus übertragene computerbasierte Test? Wenn beispielsweise für den computerbasierten Test andere Antwortformate (bei gleichem Aufgabeninhalt) verwendet wurden, dann lässt sich ohne dieses Wissen auch die Bedeutsamkeit der gefundenen Unterschiede schlecht beurteilen.

Eine weitere wichtige Frage ist, welche Kriterien zur Beurteilung von Äquivalenz in den einzelnen Studien herangezogen und mit welchen statistischen Analyseverfahren diese untersucht wurden. Dabei sind auch immer die Aussagekraft dieser Verfahren und die Power der Studien zu beachten. So bieten latente Verfahren Vorteile gegenüber manifesten Verfahren, Inner-Subjekt-Designs haben eine höhere statistische Power als Zwischen-Subjekt-Designs usw. Die Kriterien einer Äquivalenzuntersuchung sollten klar definiert und systematisch belegt werden, damit die Unterschiede zwischen verschiedenen Studien auch nach diesen beurteilt werden können. So können beispielsweise Studien, die sich nur auf deskriptive Vergleiche zwischen Testscores beziehen, nicht auf einer Ebene verglichen werden mit solchen, welche Äquivalenz auf unterschiedlichen Ebenen und in Bezug auf verschiedene psychometrische Eigenschaften beurteilen. In der vorliegenden Dissertation ist es gelungen, Äquivalenzkriterien herauszuarbeiten und anschaulich mit Analysemethoden zu verbinden, die geeignet sind, Modus-Effekte zu untersuchen. Die Voraussetzungen für die Anwendung dieser Methoden, die zum Teil im Studiendesign, aber auch in der Bestimmung eines geeigneten Skalierungsmodells liegen, wurden unter Berücksichtigung verschiedener Aspekte des Studiendesigns und im Hinblick auf ihre Vor- und Nachteile diskutiert. Die Äquivalenzkriterien, namentlich das erfasste Konstrukt, die Reliabilität sowie die Item-Diskrimination und die Item-Schwierigkeit wurden in ihrer Bedeutung für eine Äquivalenz-Studie ausführlich dargestellt. Bei der konkreten Beschreibung der Vorgehensweise einer Modus-Effekt-Studie und dem Aufzeigen methodischer Wege, Äquivalenz zu untersuchen, wurde der Modus-Effekt als eigener Parameter eingeführt. Dieser ermöglicht die Abbildung von Modus-Effekten, beispielsweise in der Schwierigkeit der Items. Der erste Beitrag beschreibt dies als sukzessive Vorgehensweise, indem die verschiedenen Schritte und Ebenen einer Modus-Effekt-Studie anhand der oben genannten Kriterien detailliert dargestellt und diskutiert werden. Dieser Beitrag kann somit als schematische Handlungsanleitung für eine Äquivalenzuntersuchung dienen. Damit begegnet er der oben genannten Kritik einer fehlenden Systematik von Modus-Effekt-Studien und trägt wesentlich dazu bei, dass, unter Berücksichtigung der beschriebenen Aspekte, zukünftige Studien systematischer und vor allem auch vergleichbarer werden können. Der Beitrag kann zudem bereits in der Studienplanung dienlich sein, wenn es um die Wahl des geeigneten Designs geht und um die Fragen, ob der Modus beispielsweise zwischen oder innerhalb der Personen variiert werden soll und welche Vor- oder Nachteile dies in Hinblick auf die spätere Auswertung mit sich bringt.

Als wichtigstes Äquivalenzkriterium wird in der vorliegenden Arbeit die Konstrukt-Äquivalenz besprochen. Trotz seiner Bedeutsamkeit wurde es in vielen Studien nicht oder nur mit Methoden eingeschränkter Aussagekraft (z. B. mittels manifester Korrelation der Testscores) untersucht. Dabei ist gerade im Hinblick auf die Validität von Testscore-Interpretationen von ausschlaggebender Bedeutung, dass sich das Konstrukt durch einen Moduswechsel nicht ändert. Denn nur wenn das erfasste Konstrukt beim Wechsel des Administrationsmodus unverändert bleibt, können Ergebnisse auch dann verglichen und Testscores eines computerbasierten Tests analog zu denen eines papierbasierten Tests interpretiert werden, wenn sich beispielsweise die Schwierigkeit des gesamten Tests oder aber nur einzelner Items verändert. Vergleiche zwischen Testscores aus unterschiedlichen Modi sind somit nur in Verbindung mit dem Nachweis von Konstrukt-Äquivalenz zulässig, denn dann lässt sich eine gemeinsame Metrik für Testscores aus PBA und CBA bilden. Die empirischen Analysen, die in dieser Dissertation entlang der oben beschriebenen Kriterien durchgeführt wurden, zeigten anhand der Daten dreier Leseverständnistests Evidenz dafür, dass das Konstrukt zwischen den Modi äquivalent war (siehe die Beiträge II und III). Die gefundenen Schwierigkeitsunterschiede zwischen den Items, konnten somit über modusspezifische Item-Parameter (Freisetzung des jeweiligen Parameters) zwischen den Modi adjustiert und eine gemeinsame Metrik für die Interpretation der Testscores hergestellt werden. Einschränkend ist dabei jedoch anzumerken, dass zumindest in der einen Studie Konstrukt-Äquivalenz über die Beziehung zu anderen Variablen überprüft werden musste, da nur ein Zwischen-Subjekt-Design vorlag. Der Berechnung von Cross-Mode-Korrelationen, die nur aus Daten eines Inner-Subjekt-Designs bestimmt werden können (vgl. Mead & Drasgow, 1993), ist eine höhere statistische Power zuzusprechen. Sie wurde in der im Beitrag II beschriebenen Studie angewendet.

Ist zwar das Konstrukt äquivalent, aber lassen sich Modus-Effekte in Bezug auf die Schwierigkeit einzelner Items finden, dann ist bei der Beurteilung der Bedeutsamkeit dieser Effekte auch immer zu berücksichtigen, wie stark die Computerisierung von der Gestaltung der Aufgaben im Testheft abweicht. Denn selbst wenn die Aufgaben innerhalb der Tests identisch sind, so unterscheiden sich papierbasierte Tests doch in einigen äußeren Kriterien von computerbasierten Tests. Finden sich signifikante Schwierigkeitsabweichungen für eine Aufgabe zwischen Computer und Testheft, so sollte man sich die Eigenschaften dieser Aufgabe ansehen. Dabei kann sich das Layout, aber auch die Handhabbarkeit der Aufgabe zwischen den Modi erheblich unterscheiden. Die Erforschung von schwierigkeitsbestimmenden Eigenschaften computerisierter Testaufgaben in Verbindung mit

ihrem Einfluss auf Modus-Effekte trägt wesentlich zur Optimierung der Gestaltung von Tests am Computer bei. Die vorliegende Arbeit liefert einen methodischen Ansatz dazu, wie Item-Eigenschaften systematisch in die Analyse einbezogen werden und ihre Effekte auf Modusunterschiede untersucht werden können. Als Eigenschaften können dabei alle Merkmale herangezogen werden, die der Klassifikation der Items bzw. der jeweiligen Administrationsmodi dienen. In der empirischen Untersuchung zweier Leseverständnistests (beschrieben im dritten Beitrag) stellte sich dabei vor allem das Antwortformat der sogenannten Combo-Box als schwierigkeitsverändernd heraus. Zuordnungsaufgaben, die im Testheft durch das Eintragen eines Buchstabens bearbeitet wurden, zeigten am Computer eine erhöhte Schwierigkeit, wenn dieser Buchstabe aus einem Drop-Down-Menü ausgewählt werden musste. Dieses Ergebnis kann dazu beitragen, dass alternative Umsetzungsmöglichkeiten für Zuordnungsaufgaben entwickelt und in zukünftigen Studien erprobt werden. Eine Möglichkeit für die computerisierte Umsetzung wäre beispielsweise der Einsatz der Funktion „Drag and Drop“. Dies wäre stärker an die Bedienung von Touch-Geräten (Smartphones, Tablets) angelehnt, was eine größere Vertrautheit der Testpersonen mit dieser Bearbeitungsform erwarten ließe. Aber auch diese Umsetzungsmöglichkeit sollte in systematisch angelegten Modus-Effekt-Studien untersucht werden, um einen Effekt auf die Item-Schwierigkeit auszuschließen, insbesondere dann, wenn die Bedienung über die Computer-Maus erfolgen soll. Neben dem oben beschriebenen Antwortformat wurde in der Analyse außerdem eine erhöhte Schwierigkeit für Items an erster und zweiter Position einer Unit am Computer gefunden. Die ersten beiden Items einer Unit wurden im Testheft auf einer Doppelseite zusammen mit dem Lesetext präsentiert, während dies am Computer nicht möglich war und die Navigation zwischen Text und Items mittels „Weiter“- und „Zurück“-Button erfolgen musste. Aus der erhöhten Schwierigkeit dieser Items am Computer gegenüber denselben Items im Testheft lässt sich möglicherweise schließen, dass sich Aufgaben dann leichter bearbeiten lassen, wenn sie direkt auf derselben Seite wie der dazugehörige Text angeboten werden. Auch diese Annahme lässt sich für die Erforschung der optimalen Gestaltung von (Computer-) Tests heranziehen und prüfen. Den Lesetext auf einer Seite mit den dazugehörigen Items zu präsentieren, ist am Computer aufgrund der gegenüber dem DIN-A4-Papierformat geringeren Größe des Bildschirms zumeist nicht möglich. Hier könnte aber überlegt werden, den Text aufzuteilen und somit jeweils nur relevante Ausschnitte mit den jeweiligen Items gemeinsam zu präsentieren, das heißt alle für ein Item relevanten Informationen auf einen Blick darzubieten (vgl. Pommerich & Burden, 2000). Soll ein papierbasierter Test gleichzeitig mit einem computerbasierten Test vorgelegt und Äquivalenz

gewährleistet werden, ist es jedoch natürlich wichtig, in beiden Versionen dieselbe Ansicht des Textes zu verwenden. Auch ist denkbar, dass die Testteilnehmer dieser Studie von der Möglichkeit Gebrauch gemacht haben, ihnen wichtig erscheinende Textpassagen im Testheft zu unterstreichen (vgl. Pommerich, 2004). Das visuelle Hervorheben von Schlüsselwörtern oder Textpassagen könnte der mentalen Speicherung der visuell-räumlichen Darstellung des Textes dienlich gewesen sein. Dabei wird das Einprägen der Position bestimmter Wörter im Text für Lesetexte am Computer im Vergleich zum Testheft als schwieriger eingeschätzt. Dies wird im Zusammenhang mit Modus-Effekten ebenfalls als ein möglicher Grund für eine erhöhte Schwierigkeit von Lesetests am Computer diskutiert (Kerr & Symons, 2006; Mangen, Walgermo, & Brønnick, 2013). In dieser Studie könnten somit Markierungen innerhalb des Textes vor allem bei der Bearbeitung der auf derselben Seite dargebotenen ersten beiden Items nützlich gewesen sein, da die hervorgehobenen Informationen unmittelbar beim Bearbeiten der Items sichtbar waren. Das (farbliche) Hervorheben bzw. Unterstreichen von Wörtern oder Textteilen ist inzwischen auch bei der Computerisierung von Testaufgaben umsetzbar (z. B. mittels CBA-Itembuilder, Rölke, 2012). Daher kann ihre Übertragung auf den Computer ebenfalls dazu beitragen, Schwierigkeitsveränderungen zu reduzieren (vgl. Kingston, 2009).

Ein weiterer wesentlicher Aspekt beim Übergang von papier- zu computerbasiertem Assessment ist das Gütekriterium der Testfairness. Per Definition dürfen Testwerte nicht zu einer systematischen Benachteiligung von Personen aufgrund ihrer personenbezogenen Merkmale (z. B. des Geschlechts) führen bzw. Tests bestimmte Personen nicht gegenüber anderen Personen mit derselben (wahren) Fähigkeit bevor- oder benachteiligen (z. B. Kubinger, 2009). Im Hinblick auf den Wechsel des Administrationsmodus‘ vom PBA zum CBA heißt dies, dass das computerbasierte Testen zu keiner systematischen Benachteiligung von Personen, beispielsweise aufgrund ihrer Computerfähigkeiten, ihrer Erfahrungen mit dem Computer oder auch ihres Geschlechts, führen darf (vgl. Jurecka & Hartig, 2007). Diese Dissertation zieht dazu relevante Persönlichkeitsvariablen, nämlich die Computerfähigkeit und das Geschlecht einer Person, zur Erklärung von differentiellen Effekten des Computers als Administrationsmodus‘ heran. Im Ergebnis zeigte sich keine systematische Benachteiligung durch das computerbasierte Testen aufgrund der Computerfähigkeit oder des Geschlechts. Auch dieses Ergebnis kann als eine Evidenzquelle für Konstrukt-Äquivalenz interpretiert werden, da die Einführung von computerbasierten Tests keine zusätzliche konstrukt-irrelevante Varianz bedingt durch Unterschiede in der Computerfähigkeit hervorbrachte (vgl. Russel et al., 2003).

Die Ergebnisse der vorliegenden Dissertation unterstreichen nochmals die Notwendigkeit von Äquivalenzstudien und betonen ihre Bedeutsamkeit. Neben der Durchführung von Modus-Effekt-Studien im Einzelfall scheint es dabei jedoch genauso wichtig, eben nicht nur die jeweilige Einzelstudie hinsichtlich des eigensetzten Tests in den Blick zu nehmen, sondern auch die Gesamtheit aller Studien mit dem Ziel, generalisierbare Indikatoren zu gewinnen, nicht aus den Augen zu verlieren. Dabei ist von ausschlaggebender Bedeutung, verbindliche Kriterien für die Untersuchung von Modus-Effekten sowie geeignete statistische Verfahren als Standards festzulegen. Dies ermöglicht es, ein einheitlicheres Bild von Modus-Effekten zu generieren und überdies wissenschaftlich fundierte Indikatoren abzuleiten, welche der erfolgreichen Erstellung von computerbasierten Tests dienen. Das Ziel bzw. die selbstgestellte Aufgabe einer jeden Äquivalenz-Studie sollte es also auch sein, diejenigen Eigenschaften zu identifizieren, die verstärkt zu Modus-Effekten beitragen. Denn nur so ist es möglich, diejenigen Aspekte der Testerstellung zu identifizieren, durch welche sich die auftretenden Unterschiede erklären lassen, und sie bei zukünftigen Computerisierungen zu verbessern. Natürlich wird es immer auch zufällige Modus-Effekte geben. Wenn aber jede einzelne Studie dazu beizutragen vermag, systematische Effekte durch eine entsprechende Gestaltung der Computer-Tests zukünftig zu minimieren, dann hat die Wissenschaft mit ihrer Vielzahl von Modus-Effekt-Studien einen bedeutsamen Beitrag zum Wechsel von papier- auf computerbasiertes Assessment geleistet.

6. Ausblick: Modus-Effekt-Studien – Relevanz bis morgen. Die Nutzung neuer Technologien in Assessments

Gegen die Relevanz zukünftiger Modus-Effekt-Studien könnten kritische Leser nun einwenden, inzwischen würden ohnehin alle Testungen und Befragungen am Computer durchgeführt und neuere Tests nur noch für den Computer entwickelt und somit werde auch ein Vergleich mit Papier-Tests obsolet. Um diesen Einwand zu entkräften, bedarf es nur eines Blickes auf die rasante Weiterentwicklung und den Einsatz neuer Technologien für Assessments. Die Frage von Modus-Effekten stellt sich nämlich nicht nur im aktuellen Kontext des Moduswechsels vom Papier zum Computer in den „großen“ Studien. Testungen an Smartphones und Tablets ergänzen schon heute (z. B. Huff, 2015; Ling, 2015) die Möglichkeiten von Assessments am Computer. Dabei spielt auch die Nutzung dieser technischen Geräte im Alltag eine große Rolle. So zeigte zum Beispiel eine Umfrage von

Pearson (Pearson, 2015), dass knapp 90 Prozent der befragten College-Studenten Laptops und Smartphones in ihrem schulischen und außerschulischen Alltag nutzen. Beobachtet man sein eigenes Umfeld, dann scheint sich dies zu bestätigen. Es mutet an, dass die meisten Kinder, Jugendlichen und Erwachsenen in ihrem Alltag inzwischen kaum noch auf ihr Smartphone sowie die davon ausgehende Mobilität verzichten können. Um Testteilnehmer weiterhin erreichen zu können, wird es also notwendig sein, Testungen an die Schnelllebigkeit und die Flexibilität des Alltags anzupassen und in diese zu integrieren. Vorhandene papierbasierte, aber auch computerbasierte Tests müssen sodann auf andere Hardware übertragen werden. Dies erfordert auch zukünftig die Erforschung von Äquivalenz, dann vielleicht weniger zwischen papier- und computerbasierten Tests, sondern zwischen Tests auf verschiedenen Arten von Hardware, etwa zwischen Tablets (mit oder ohne externer Tastatur) und Smartphones (Illingworth et al., 2015), und in verschiedenen Settings (Ihme et al., 2009). Die Umsetzung von Papier-Tests auf den Computer erscheint dabei noch vergleichsweise einfach, ist doch der Computerbildschirm in seiner Größe den DIN-A4-Testbögen noch weitgehend ähnlich. Sollen nun aber Smartphones genutzt bzw. zusätzlich als (wählbarer) Testmodus angeboten werden, so ist die Eins-zu-Eins-Übertragung eines Tests gar nicht mehr möglich (vgl. Hancock, Sawyer, & Stafford, 2015; Huff, 2015; Ling, 2016). Hier sind Fragen wie die folgenden zu klären: Soll auf dem Smartphone dieselbe Testseite wie auf dem Computer dargestellt werden, der Testteilnehmer hätte damit allerdings zu scrollen oder zu zoomen? Oder sollen – insbesondere bei Tests zum Leseverständnis mit langen Lesetexten ist dies relevant – weniger Informationen auf einer Seite präsentiert werden, sodass der Testteilnehmer häufiger blättern bzw. „wischen“ muss? Damit stünde beispielsweise auch die Frage im Raum, ob die benötigte Lese-Zeit dadurch erhöht würde und wie vieler Navigationsrestriktionen es bedarf, um eine durch die größere Anzahl von Testseiten herbeigeführte, „virtuelle Verirrung“ innerhalb des Tests seitens der Teilnehmer zu verhindern. Die Gestaltung der Übertragung von Computer-Tests auf andere technische Geräte wirft somit neue Fragestellungen hinsichtlich der Äquivalenz auf. Gleichermaßen bedeutsam bleiben dabei selbstverständlich die oben genannten Äquivalenzkriterien, allen voran das der Konstrukt-Äquivalenz. Und auch hier gilt, dass gerade die aufgrund der neuen Gestaltungsmöglichkeiten veränderten Test- und Item-Eigenschaften einer sorgfältigen Erforschung dahingehend bedürfen, inwiefern sie einen Effekt auf die Schwierigkeit haben. Auch im Hinblick auf die Testfairness wird es weiterhin wichtig sein, differentielle Effekte des Administrationsmodus’ zu berücksichtigen. Allerdings ist zu vermuten, dass in Zukunft zumindest für die Computerfähigkeiten, aber möglicherweise auch im Hinblick auf

Geschlechtsunterschiede weniger signifikante Effekte auftreten, da Kinder beiderlei Geschlechts bereits immer früher mit den neuen Techniken vertraut werden.

Neben der Herausforderung der Äquivalenz zwischen den Technologien der neuen und denen der älteren Generation (wozu in ein paar Jahren wahrscheinlich auch Computer gehören könnten), ergeben sich auch weitere, die es zu bewältigen gilt. Dazu gehören die nicht kontrollierbaren Bedingungen von Online-Test-Settings sowie die Herausforderungen des Datenzugriffs und des Datenschutzes. Online-Testungen können eine sinnvolle Ergänzung zu Assessments in herkömmlichen Gruppentest-Settings darstellen (vgl. Frein, 2011; Ihme et al., 2009). Durch die Wahl von Ort und Zeit der Test-Bearbeitung können Testteilnehmer diese flexibler in ihren Alltag integrieren (Frein, 2011; Jurecka & Hartig, 2007). Zudem lassen sich durch Online-Tests solche Personen leichter erreichen, die aufgrund ihrer eingeschränkten Mobilität bislang nur wenige Möglichkeiten dazu hatten, an Studien teilzunehmen, was wiederum die Repräsentativität der durchgeführten Studien erhöhen kann (Noyes & Garland, 2008). Die unterschiedlichen Test-Settings, wie das herkömmlicherweise vor allem in LSA verwendete Gruppentest-Setting gegenüber dem, durch den Einsatz neuer Technologien in Assessments hinzukommende Online-Test-Setting, sind also ebenfalls wesentliche Aspekte von Äquivalenz-Fragestellungen (vgl. Frein, 2011). Eine wichtige Frage dabei ist die der Standardisierung. Während Testungen in Gruppentest-Settings durch kontrollierte Umgebungsbedingungen und die Anwesenheit von Testleitern noch (mehr oder weniger) standardisiert gehalten werden können, gibt es bei Online-Testungen, deren Zeit und Ort selbst gewählt werden können, keine Möglichkeit, unkontrollierbare Störeffekte auszuschalten bzw. Testbedingungen vergleichbar zu halten (Ihme et al., 2009; Jurecka & Hartig, 2007). Hinweise wie solche, die Testung selbstständig und nur an einem ruhigen Ort und ohne Unterbrechung vorzunehmen, können dabei zwar hilfreich sein und im Einzelfall befolgt werden, entziehen sich jedoch jeglicher Kontrolle des Auftraggebers. Auch Fragen zur Bearbeitungssituation im Anschluss an die Testung können Informationen liefern, sofern sie ehrlich beantwortet wurden. Hier stellt sich dann aber wiederum die Frage, wie mit solchen Daten umgegangen werden soll. Sind Testteilnehmer, die sich zum Beispiel auf dem Nachhauseweg im Zug des Tests angenommen und abschließende Fragen zu den Bearbeitungsbedingungen ehrlich beantwortet haben, auszuschließen, weil die Bedingungen im Zug als nicht optimal anzusehen sind? Auch hier ist die Äquivalenzforschung (zum Beispiel auch unter Zuhilfenahme von Log-Daten zur Auswertung von Reaktionszeiten) notwendig, um Antworten auf diese Fragen zu finden. Um Testbedingungen möglichst vergleichbar zu halten, ist es überdies wichtig, den Testzugriff zu kontrollieren. So sollte

beispielsweise über Zugangscodes versucht werden mehrfache Testzugriffe zu verhindern (vgl. Cronk & West, 2002; Jurecka & Hartig, 2007; Noyes & Garland, 2008), was zusätzlich dem Schutz des Testmaterials dienlich ist. Ein weiterer Aspekt, der im Zusammenhang mit selbstgewählten Test-Settings bei der Erstellung von Online-Tests diskutiert werden muss, ist die Zuhilfenahme von externen Quellen bei der Testbearbeitung. Während Testleiter auch gleichzeitig dafür Sorge tragen können, dass keine Bearbeitungshilfen (Suchmaschinen auf Smartphones, Lehrbücher usw.) verwendet werden, so ist dies im Kontext von Online-Test-Settings nur erschwert möglich. Zeitbeschränkungen für einzelne Items sowie den gesamten Test können als Möglichkeiten in Erwägung gezogen werden, die Verwendung externer Hilfen zu verhindern (vgl. Frein, 2011). Aber auch bei einer solchen Implementierung müssen potenzielle, daraus resultierende Modus-Effekte ausgeschlossen werden, um Äquivalenz zu bestehenden Assessments (PBA oder CBA) zu gewährleisten und die Ergebnisse vergleichen zu können. Dieser abschließende Abschnitt macht deutlich, dass die Relevanz von Modus-Effekt-Studien also keineswegs mit der gründlichen Erforschung der Äquivalenz von papier- und computerbasiertem Testen endet, sondern auch in Zukunft bestehen bleibt, genährt durch die rasante technologische Weiterentwicklung und die immerwährende Herausforderung für die psychologische Diagnostik, möglichst standardisierte Testbedingungen zu schaffen.

Literaturverzeichnis

Alexander, M. W., Bartlett, J. E., Truell, A. D., & Ouwenga, K. (2001). Testing in a Computer Technology Course: An Investigation of Equaivalency in Performance Between Online and Paper and Pencil Methods. *Journal of Career and Technical Education*, *18 (1)*, 69-80.

American Educational Research Association [AERA], American Psychological Association (APA), & National Council on Measurement in Education [NCME]. (2014). *Standards for Educational and Psychological Testing*. Washington: AERA, APA, NCME.

Artelt, C., Weinert, S., & Carstensen, C. H. (2013). Assessing competencies across the lifespan within the German National Educational Panel Study [NEPS] – Editorial. *Journal for educational research online, 5* (2), 5-14.

Association of Test Publishers [ATP]. (2000). *Guidelines for computer-based testing*. Washington DC: Association of Test Publishers.

Bennett, R. E. (2003). Online Assessment and the Comparability of Score Meaning. *Research Memorandum 3* (5), 1-19.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. *The Journal of Technology, Learning, and Assessment, 6* (9).

Bennett, R. E., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing complex problem-solving performances. *Assessment in Education, 10*, 347-359.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examiner's ability. In F. M. Lord, & M. R. Novick (Hrsg.), *Statistical theories of mental test scores* (S.17-20). Reading: Addison-Wesley.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research, 31* (1), 51-60.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance.* (ETS-RR-01-23), Princeton, NJ: Educational Testing Service.

Choi, S. W., & Tinkler, T. (2002, April). Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting. Paper presented at the annual meeting of

the National Council on Measurement in Education, New Orleans, LA.

Clariana, R., & Wallace, P. (2002). Paper–based versus computer–based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593-602.

Cronk, B.C., & West, J.L. (2002). Personality research on the Internet: A comparison of Web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments, & Computers, 34* (2), 177-180.

Davey, T. (2011). Practical Considerations in Computer-Based Testing. Verfügbar unter http://www.ets.org/Media/Research/pdf/CBT-2011.pdf

Educational Testing Service (ETS). (2015). *PISA 2015 Field trial analysis report: Outcomes of the cognitive assessment* [interner Bericht]. Princeton, NJ.

Frein, S. (2011). Comparing in-class and out-of-class computer-based tests to traditional paper-and-pencil-tests in introductory psychology courses. *Teaching of Psychology, 38* (4), 282-287.

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. Sonderh., 16* (1) , 53-57.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for educational research online, 5* (2), 50-79.

Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing Individual differences in Basic Computer Skills: Psychometric characteristics of an interactive performance measure. *European Journal of Psychological Assessment, 29*, 263-275.

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based assessment. In D. Leutner, J. Fleischer, J. Grünkorn & E. Klieme (Hrsg.), *Competence assessment in education: Research, models and instruments* (S. 407-425). Heidelberg: Springer.

Gould, J. D., & Grischkowsky, N. (1984). Doing the same work with hard copy and with cathode-ray tube (CRT) computer terminals. *Human Factors, 26* (3), 323-337.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical Guidelines for Assessing Computerized Adaptive Tests. *Journal of Educational Measurement, 21* (4), 347-360.

Hancock, P.A., Sawyer, B.D., & Stafford, S. (2015). The effects of display size on performance. *Ergonomics, 58* (3), 337-354.

Hartig, J., Kröhne, U., & Jurecka, A. (2007). Anforderungen an computer- und netzwerkbasierte Assessments. In J. Hartig, & E. Klieme, (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (S. 57-67). Berlin: BMBF.

Heerwegh, D., & Loosveldt, G. (2002). An Evaluation of the Effect of Response Formats on Data Quality in Web Surveys. *Social Science Computer Review, 20* (4), 471-484.

Heine, J.-H., Mang, J., Borchert, L., Gomolka, J., Kröhne, U., Goldhammer, F., & Sälzer, C. (2016). Kompetenzmessung in PISA 2015. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme & O. Köller (Hrsg.), *PISA 2015: Eine Studie zwischen Kontinuität und Innovation* (S.383-430). Münster: Waxmann.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the Effect of Computer-Based Passage Presentation of Reading Test Performance. *The Journal of Technology, Learning, and Assessment, 3* (4), 1-35.

Holland, P. W., & Dorans, N. J. (2006). Linking and Equating. In R. L. Brennan (Hrsg.), *Educational measurement* (4. Aufl., S. 187-220). Westport, CT: Praeger.

Hox, J. J., de Leeuw, E. D., & Zijlmans, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology, 6*.

Huff, K. C. (2015). The comparison of mobile devices to computers for web-based assessments. *Computers in Human Behavior, 49*, 208-212.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practices, 20* (3), 16-25.

Ihme, J. M., Lemke, F., Lieder, K., Martin, F., Müller, J. C., & Schmidt, S. (2009). Comparison of ability tests administered online and in the laboratory. *Behavior Research Methods, 41* (4), 1183-1189.

Illingworth, A. J., Morelli, N. A., Scott, J. C., & Boyd, S. L. (2015). Internet-Based, Unproctored Assessments on Mobile and Non-Mobile Devices: Usage, Measurement Equivalence, and Outcomes. *Journal of Business and Psychology, 30* (2), 325-343.

International Test Commission (ITC). (2005). *International Guidelines on Computer-Based and Internet Delivered Testing*.

Jurecka, A., & Hartig, J. (2007). Computer- und netzwerkbasiertes Assessment. In J. Hartig, & E. Klieme, (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (S. 37-48). Berlin: BMBF.

Kerr, M.A., & Symons, S.E. (2006). Computerized Presentation of Text: Effects on Children's Reading of Informational Material. Reading and Writing, 19 (1), 1-19.

Kiefer, T., Robitzsch, A., & Wu, M. (2016). *TAM: Test analysis modules*. (R package version 1.15-0).

Kim, D.-H., & Huynh, H. (2008). Computer-Based and Paper-and-Pencil Administration Mode Effects on a Statewide End-of-Course English Test. *Educational and Psychological Measurement, 68* (4), 554-570.

Kingston, N. M. (2009). Comparability of Computer- and Paper-Administered Multiple-Choice Tests for K–12 Populations: A Synthesis. *Applied Measurement in Education, 22,* 22-37.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2. Auflage). New York: Springer.

Kubinger, K. D. (2009). *Psychologische Diagnostik. Theorie und Praxis psychologischen Diagnostizierens* (2. Auflage). Göttingen: Hogrefe.

Kröhne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft, 14* (2), 169-186.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing, 6,* 1-24.

Ling, G. (2016). Does It Matter Whether One Takes a Test on an iPad or a Desktop Computer? *International Journal of Testing, 16* (4), 352-377.

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. International Journal of Educational Research, 58, 61-68.

Mazzeo, J., & Harvey, A. L. (1988). *The Equivalence of Scores from Automated and Conventional Educational and Psychological Tests. A Review of the Literature* (College Board Report, S. 88-8). New York: College Entrance Examination Board.

Mead, A. D., & Drasgow, F. (1993). Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. *Psychological Bulletin, 114* (3), 449-458.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *ETS Research Report Series,* (1), 1-30.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Neuman, G., Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement, 22*(1), 71-83.

Nikou, S. A., & Economides, A. A. (2016). The impact of paper-based, computer-based and mobile-based self-assessment on students' science motivation and achievement. *Computers in Human Behavior, 55* (Part B), 1241-1248.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51* (9), 1352-1375.

OECD. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. OECD Publishing.

OECD. (2010). *PISA 2009 Results: Executive Summary*. OECD Publishing.

OECD. (2009). *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science.* OECD Publishing.

Osterlind, S. J., & Everson, H. T. (2009). *Differential Item Functioning*. Thousand Oaks: SAGE Publications, Inc.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2010). Innovative Items for Computerized Testing. In W. J. van der Linden, & C. A. W. Glas (Hrsg.), *Elements of Adaptive Testing* (S. 215-230). New York: Springer.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Pearson (2015). *Pearson student mobile device survey 2015*. Verfügbar unter http://www.pearsoned.com/wp-content/uploads/2015-Pearson-Student-Mobile-Device-Survey-College.pdf

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao, & S. Sinharay (Hrsg.), *Handbook of Statistics: Vol. 26. Psychometrics,* (S. 125-167). New York: Elsevier.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large

Scale State Assessment Program. *The Journal of Technology, Learning, and Assessment*,

3 (6).

Pommerich, M. (2004). Developing Computerized Versions of Paper-and-Pencil Tests: Mode Effects for Passage-Based Tests. *The Journal of Technology, Learning, and Assessment, 2* (6), 3-44.

Pommerich, M., & Burden, T. (2000, April). *From Simulation to Application: Examinees React to Computerized Testing.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

Pomplun, M., & Custer, M. (2005). The Score Comparability of Computerized and Paper-and-Pencil Formats for K-3 Reading Tests. *Journal of Educational Computing Research, 32* (2), 153-166.

Pomplun, M., Frey, S., & Becker, D. F. (2002). The Score Equivalence of Paper-and-Pencil and Computerized Versions of a Speeded Test of Reading Comprehension. *Educational and Psychological Measurement, 62,* 337-354.

Puhan, G., Boughton, K., & Kim, S. (2007). Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. *The Journal of Technology, Learning, and Assessment, 6* (3).

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87* (3), 517-529.

Rasch, G. (1960). *Probabilistic Models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114* (3), 552-566.

Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J.-H. (2016). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. *Diagnostica 63,* 148-165.

Rölke, H. (2012). The ItemBuilder: A Graphical Authoring System for Complex Item Development. *In World Conference on E-Learning in Corporate, Gouvernmet, Healthcare, and Higher Education*. (S. 344-353), Association for the Advancement of Computers in Education (AACE).

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2., vollständig überarbeitete und erweiterte Auflage). Bern: Huber.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-Based Testing and Validity: A Look Back and Into the Future. *Assessment in Education, 10* (3), 279-293.

Sälzer, C., & Reiss, K. (2016). PISA 2015 – die aktuelle Studie. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme & O. Köller (Hrsg.), *PISA 2015: Eine Studie zwischen Kontinuität und Innovation* (S. 13-44), Münster: Waxmann.

Schröders, U., & Wilhelm, O. (2011). Equivalence of Reading and Listening Comprehension Across Test Media. *Educational and Psychological Measurement, 71* (5), 849-869.

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 (Hrsg.), *Handbook of test development* (S. 329-347), London: Lawrence Erlbaum Associates.

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models. *Journal of Educational and Behavioral Statistics, 30* (4), 443-464.

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is large. Transactions of the American Mathematical Society, 54 (3), 426-482.

Wang, S. (2004). *Online or paper: does delivery affect results? Administration mode comparability study for Stanford Diagnostic Reading and Mathematics Tests.* Pearson Education Inc., USA.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K 12 Reading Assessments: A Meta-Analysis of Testing Mode Effects. *Educational and Psychological Measurement, 68* (1), 5-24.

Wang, Y. C., Lee, C. M., Lew-Ting, C. Y., Hsiao, C. K., Chen, D.-R., & Chen, W. J. (2005). Survey of substance use among high school students in Taipei: Web-based questionnaire versus paper-and-pencil questionnaire. *Journal of Adolescent Health, 37,* 289-295.

Yamamoto, K. (2012). *Outgrowing the Mode Effect Study of Paper and Computer Based Testing.* Verfügbar unter http://www.umdcipe.org/conferences/EducationEvaluationItaly/COMPLETE_PAPERS/Yamamoto/YAMAMOTO.pdf

Anhangverzeichnis

Anhang A

Beitrag I: Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The Transition to Computer-Based Testing in Large-Scale Assessments: Investigating (Partial) Measurement Invariance between Modes. *Psychological Test and Assessment Modeling, 58* (4), 587-606.

Psychological Test and Assessment Modeling, Volume 58, 2016 (4), 597-616

The transition to computer-based testing in large-scale assessments: Investigating (partial) measurement invariance between modes

Sarah Buerger[1], *Ulf Kroehne*[1] & *Frank Goldhammer*[1,2]

Abstract

This paper provides an overview and recommendations on how to conduct a mode effect study in large-scale assessments by addressing criteria of equivalence between paper-based and computer-based tests. These criteria are selected according to the intended use of test scores and test score interpretations. A mode effect study can be implemented using experimental designs. The major benefit of combining experimental design considerations with the IRT methodology of mode effects is the possibility to investigate partial measurement invariance. This allows test scores from different modes to be used interchangeably and means of latent variables or mean differences and correlations to be compared on the population level even if some items differ in difficulty between modes. For this purpose, a multiple-group IRT model approach for analyzing mode effects on the test and item levels is presented. Instances where partial measurement invariance suffices to combine item parameters into one metric are reviewed in this paper. Furthermore, relevant study design requirements and potential sources of mode effects are discussed. Finally, an extension of the modelling approach to explain mode effects by means of item properties such as response format is presented.

Keywords: mode effect, equivalence, computer-based assessment, partial measurement invariance, anchor items

[1] *Correspondence concerning this article should be addressed to:* Sarah Buerger, PhD, German Institute for International Educational Research (DIPF), Solmsstraße 73-75, 60486 Frankfurt am Main, Germany; email: buerger@dipf.de

[2] Centre for International Student Assessment (ZIB)

If different versions of a test are used and test scores are meant to be comparable between examinees, an investigation of the differences between those test versions is required (Parshall, Spray, Kalohn, & Davey, 2002; Wang & Kolen, 2001). Currently, a frequent case in which two versions of test instruments come into play in large-scale assessments involves a change in the administration mode from paper-based (PBA) to computer-based assessment (CBA; e.g., PIAAC, OECD, 2015; PISA, OECD, 2014a). Potential differences between scores on tests administered in different modes might be the result of mode effects, even if the computerization was conducted in such a way so as to make the test versions in each mode as comparable as possible. As a result, individuals with the same ability level completing the test in different modes may not obtain the same test score, meaning that scores cannot be used interchangeably (Raju, Laffitte, & Byrne, 2002; Van den Noortgate & De Boeck, 2005). Paper-based and computer-based assessments differ in measurement properties (e.g., test layout, navigation of the test and the handling of input devices, see Kroehne & Martens, 2011), which may in turn affect the comparability of their psychometric properties. Thus, differences between modes might not only be caused by a single measurement property but also by an amalgamation of the properties described as potential sources of mode effects in the next section. The probability of mode effects is assumed to increase "the more complicated it is to present or take the test on computer" (Pommerich, 2004, pp.3-4).

A change in administration mode can have an effect on psychometric properties like construct validity and the difficulty and discrimination of the test or single items (Mead & Drasgow, 1993; Puhan, Boughton, & Kim, 2007). The intended use of test scores and test score interpretations determine which psychometric properties have to be equivalent across modes. Those properties serve as criteria of equivalence, and the goal of a mode effect study is to try to falsify their equivalence.

Reviews of the literature reveal inconsistent findings regarding the equivalence of computer- and paper-based tests, meaning that some studies have falsified the equivalence hypothesis, whereas others have not (see Wang, Jiao, Young, Brooks, & Olson, 2008 for an extensive overview). One reason for these heterogeneous findings may be that there are a wide variety of methods used in mode effect studies (cf. Schroeders & Wilhelm, 2011; Wang, et al., 2008). These range from approaches based on *classical test theory* (CTT) to those based on *item response theory* (IRT). In addition, differences in sample sizes and thus the power of statistical tests allow for the detection of some effects of the administration mode, while others remain hidden. Although many studies have found no significant mode effects, the heterogeneity of results indicates that, in general, the existence of mode effects cannot be ruled out, and there are no reliable computerization rules to prevent mode effects. Thus, the appropriateness of comparisons has to be investigated in equivalence studies, whose findings must be documented as required by testing standards (e.g., AERA, APA, & NCME, 2014; American Psychological Association, Committee on Professional Standards [COPS] and Committee on Psychological Tests and Assessments [CPTA], 1986; Association of Test Publishers [ATP], 2000; International Test Commission [ITC], 2005).

In the context of current international and national large-scale assessments, computer-based testing is becoming more and more common. Thus, the comparability of comput-

er-based and paper-based tests is highly important, since it is a prerequisite for comparing more recent computerized scores with scores from previous cycles or with participants or countries in the same cycle completing the traditional paper-and-pencil form (OECD, 2014a). Evidence of comparability between different administration modes is necessary to ensure conditions that enable stable trend measures in large-scale assessments (Mazzeo & von Davier, 2008). In PISA (Programme for International Student Assessment) 2015, the administration mode shifted completely from paper-based to computer-based assessment (a move which is also planned for the NAEP [National Assessment of Educational Progress] in the U.S. in 2017), although participating countries had the option of implementing PISA as a paper-based survey in the Main Study. The comparability of computer-based and paper-based items also had to be addressed in PIAAC (Programme for the International Assessment of Adult Competencies) 2012 both because some participants lacking computer skills took the paper version and also to link scores back to previous paper-based adult literacy studies such as ALL (Adult Literacy and Lifeskills Survey, OECD, 2013) and IALS (International Adult Literacy Survey, OECD, 2013). For NEPS (National Educational Panel Study; Blossfeld, Roßbach, & von Maurice, 2011) in Germany, computer-based testing was introduced in 2012 as an alternative to the paper-based assessment, and mode effect studies have become crucial in linking different modes over time points. What all of these large scale assessments have in common is that, contrary to instruments for individual diagnostics, the focus is mainly on comparing means across populations such as schools or countries, and on correlations with other performance-related variables. This intended use of test scores in large-scale assessments leads to specific equivalence criteria such as construct equivalence that should be investigated in order to ensure the required level of measurement invariance between tests administered in different modes.

In this paper, we propose a comprehensive multiple-group IRT (Item Response Theory) model approach to assess different levels of measurement invariance for categorical dependent variables. This general approach of testing hypotheses for equivalence with regard to relevant criteria also remedies shortcomings of previous studies, which have used diverse and sometimes inappropriate methods. This model is suitable for analyzing mode effects in experimental designs, where randomization is conducted, and aims to investigate partial measurement invariance, meaning that not all items have to be invariant between the modes. Its purpose is to ensure that the prerequisites for valid comparisons of results from different administration modes hold even if some or all items are unequal between modes, because mode effects can be represented by mode-specific item parameters that account for differences (e.g. in difficulty).

Sources of mode effects

Drawing on empirical evidence from previous studies, the following section presents measurement properties that might be potential sources of mode effects. With regard to the modelling approach proposed in this paper, these properties can be used to explain mode effects.

Input devices

Different input devices like a pen on paper, a mouse on the computer or the touchscreen on a tablet might interfere and interact with the experience of the test-takers in different ways (see Bennett, 2003; Parshall, Harmes, Davey, & Pashley, 2010; Schroeders & Wilhelm, 2010).

Test and item layout

The page or screen size and orientation – typically portrait on paper and landscape on the computer – differ between administration modes, which has an effect on the number of text pages as well as the size and placement of the text (e.g., column and line breaks). The presentation of multiple items on a page, as is commonly done on paper, versus one item at a time on the screen of a computer may also cause mode effects (Schroeders & Wilhelm, 2010).

Scrolling

If the amount of information on a page is larger than the screen, scrolling or paging with a mouse or touchpad to read a text and associated items is another potential source of mode effects. Scrolling has repeatedly been shown to be more difficult than paging (Bridgeman, Lennon, & Jackenthal, 2001; Higgins, Russell, & Hoffmann, 2005; Kim & Huynh, 2008; Kingston, 2009; Mazzeo & Harvey, 1988; Pearson Educational Measurement, 2005; Poggio, Glasnapp, Yang, & Poggio, 2005; Pommerich, 2004; Schwarz, Rich, & Podrabsky, 2003; Wang et al., 2008). However, findings from the large-scale assessment PIAAC suggest that the extent of scrolling had no significant impact on the difficulty of the items in the computer-based version (Yamamoto, 2012).

Item review

Item review, that is, whether the test-taker can go back to an item and change their answer, is often prohibited on the computer due to the prevention of backward navigation. On paper, navigation between items is typically not restricted (Pommerich & Burden, 2000; Vispoel, 2000). This aspect of test-taking flexibility may comprise a difference between modes (Bodmann & Robinson, 2004). However, Vispoel (2000) found in a low-stakes testing context that preventing item review did not affect average scores or psychometric properties, as only a very small percentage of test-takers used the opportunity to change their answers. In such cases, the individual benefit resulting from using item review increased with test-takers' ability level. Vispoel (2000) also showed that test-takers expressed a strong desire for item review opportunities, especially those exhibiting test anxiety.

Item response format

Computerized response formats often look only slightly different from the paper version, but show greater differences for more complex response formats such as assignment tasks or constructed responses (e.g., Heerwegh & Loosveldt, 2002; Parshall et al., 2010; Parshall et al., 2002; Sireci & Zenisky, 2006). The complexity of an item, that is, "the number and type of examinee interactions within a given task or item" (Parshall et al., 2002, p.9) plays an essential role in mode effects. Studies have shown that computerized multiple-choice items are less prone to mode effects because they are of lower item complexity (Bennett et al., 2008; Bodmann & Robinson, 2004; Parshall et al., 2002). Items requiring constructed responses have been shown to be more difficult on the computer than on paper (Bennett et al., 2008). The response format of drop-down boxes, which are often used for assignment tasks, also turns out to be more difficult when implemented on the computer (Heerwegh & Loosveldt, 2002). Mode effect studies for NEPS (Buerger, Kroehne, & Goldhammer, 2015) and PIAAC (Yamamoto, 2012) also focused on item response formats as possible sources of differences between modes. In NEPS, items with drop-down boxes on the computer showed higher difficulty (Buerger et al., 2015). The computer-based response formats used in PIAAC, such as highlighting, clicking, or scrolling, had no effect on item difficulty (Yamamoto, 2012).

Interaction of mode with test-takers' characteristics

Mode effects might be influenced by (computer-related) test-taker characteristics, which might interact with properties of the test administration and affect test-takers' performance (Kroehne & Martens, 2011). For instance, a person's general familiarity with computers has been shown to interact with the mode of administration: Students with a high degree of familiarity had an advantage in a computer-based test over students with a low degree of familiarity (Bennett et al., 2008; Clariana & Wallace, 2002; Wang et al., 2008). However, other studies have found no interaction with test-takers' computer literacy and computer use (Bennett, 2002; Higgins et al., 2005). In large-scale assessments, where populations and sub-populations are the primary focus of test score interpretations, it is crucial to consider that mode effects may vary across countries if countries differ, for instance, in the overall accessibility and usage of computers.

Design requirements and identification of groups

When comparing different modes, the assignment of persons to mode represents the main relevant aspect of the study design. The investigation of criteria (e.g. construct equivalence) for falsifying the equivalence hypothesis as well as related inferences are primarily dependent upon the design of the mode effect study (Wang et al., 2008). Therefore, when planning a mode effect study, the decision regarding mode assignment (see Figure 1) needs thorough consideration: The mode can vary between persons (between-

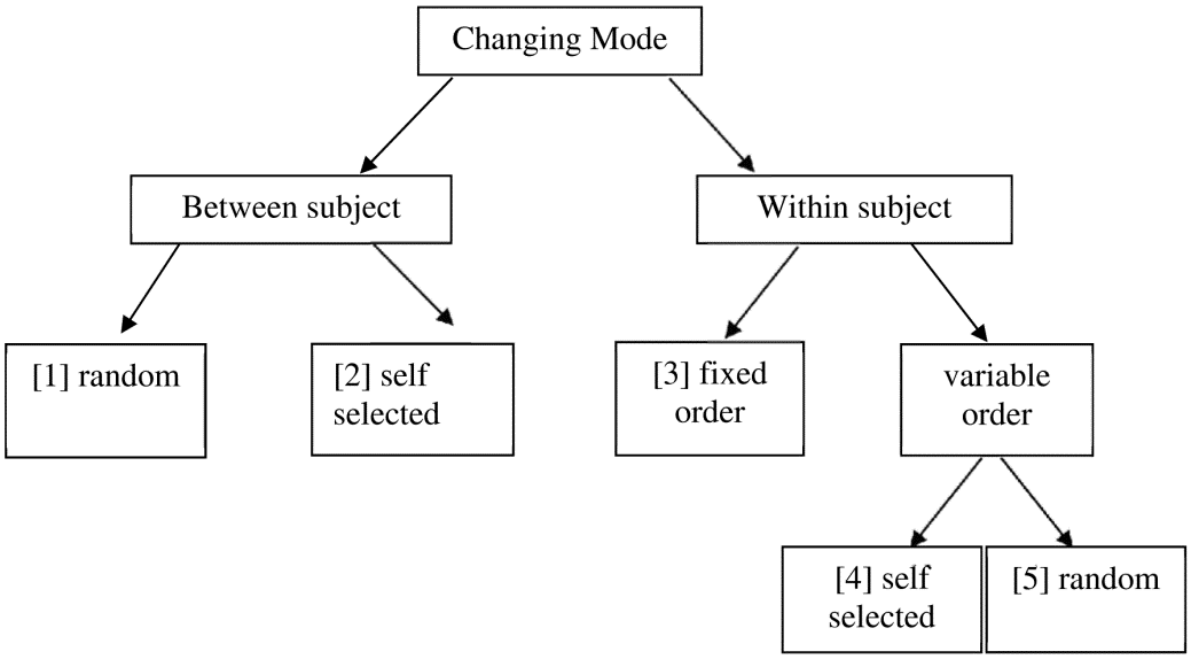


Figure 1:
Design for Mode Effect Study. Mode changing between subjects with randomization to the mode or by self-selection; or within subjects with fixed or variable order of the test parts.

subject design), meaning that each test-taker takes the test in only one mode, or it can vary within persons (within-subject design), meaning that the test is split into two parts and both parts are administered to every person in different modes. In the case of a between-subject design, there is the further distinction of whether the persons are randomly assigned to a mode [1] or get to choose the mode themselves [2] (see below). In a within-subject design, the order of the modes has to be considered. A fixed order refers to the situation when all persons take both test parts in the same order [3] (e.g., Pomplun, Frey, & Becker, 2002). With a variable test order, the question arises whether persons are allowed to choose the order in which they will take the test parts themselves [4] (e.g., Puhan et al., 2007), or whether the order is balanced and persons are randomly assigned to start with a certain mode [5] (e.g., Kim & Huynh, 2008). As the position of the test parts (modes) could have an effect, a balanced order of the test parts [5] is preferable.

The within-subject design is advantageous in several respects. It has higher statistical power because the error variance associated with individual differences is reduced and results do not depend on the assignment of persons to mode groups (Schroeders & Wilhelm, 2011). To overcome the disadvantages of a between-group design, one has to enlarge samples and ensure comparability between groups. Thus, the random assignment of test-takers to mode groups (cf. Holland & Dorans, 2006) is crucial for the interpretation of test score differences between modes. Only a design with randomly equivalent groups [1, 3, 5], ensures that the differences between modes do not occur simply due to non-random ability differences (e.g., Osterlind & Everson, 2009). Differences are then a direct result of the change in administration mode.

Groups are non-randomly equivalent if assignment to a certain mode depends, for instance, on the test-taker's decision [2, 4], meaning that essential person-related variables are not automatically balanced (Kingston, 2009). Then, comparability needs to be ensured in a different way. One possibility to make the ability distribution as comparable as possible is to use person-level covariates. In the case of self-selection of the mode, additional data on the decision-making process is necessary to build equivalent groups, for instance, with the help of matching techniques such as propensity score matching (e.g., Hox, de Leeuw, & Zijlmans, 2015). If no information about the decision-making process is available, items that can be assumed to be equal between modes are required in order to create a common ability metric.

When planning the design, how the criterion of construct equivalence is to be investigated also plays a role. The extension of the randomly equivalent group design to an order-balanced within-subject design [5] allows for the estimation of (latent) correlations between the modes and is another advantage of this design. In this case, the question of construct equivalence can be addressed by investigating whether the cross-mode correlation is not significantly different from 1. This is impossible in between-group designs, which are frequently used in large-scale assessments such as PISA, PIAAC and NEPS (Rutkowski, von Davier, & Rutkowski, 2013). Here, external criteria as suggested by a nomological network of the construct need to be measured in both groups to analyze the relationship between the test in both modes and these external criteria. For instance, for a reading comprehension test, tests of basic reading skills, such as a lexical decision task or a sentence verification task, can serve as sources of convergent evidence, while a test from a different domain (e.g. science) can be used as a source of divergent evidence. In any case, if the results from the paper-based and the computer-based assessments are equally correlated with the external criteria, the investigated hypothesis of construct equivalence is not falsified. A difference in correlations suggests that another (mode-specific) construct is also being assessed. The administration mode of criterion variables needs to be decided as well, considering that according to the *multi-trait multi-method* perspective (e.g., Eid, 2006), the correlations of tests in the same mode might be higher.

Note that an experimental mode effect study, where randomly equivalent groups complete paper and computer versions of an instrument, differs from studies investigating differential item functioning (DIF). In DIF studies, groups differ in person-level variables such as gender or mother tongue, meaning that groups cannot be randomly equivalent, whereas the instruments they complete are identical. In DIF studies, the usual way to create a common metric and thus ensure the comparability of test scores across groups is to use so-called anchor items, which are assumed to be invariant between groups and thus show no differential functioning (e.g., Reise, Widaman, & Pugh, 1993). Regarding the assessment of mode effects, items not affected by the mode change and therefore forming an anchor can be used to identify a common metric with respect to the targeted construct, allowing differences between non-anchor items to be identified as mode effects. Note that this is possible even for non-equivalent groups and not necessary for random equivalent groups, where the common metric is ensured with equal ability distributions that are guaranteed if randomization was successful.

Multiple-group IRT model for analyzing mode effects on equivalence criteria

As shown above, a broad range of methods have been used to test the existence of mode effects in previous studies. The choice of the design and the related question of whether all or at least some items should be invariant between modes is not addressed explicitly in most approaches (as for instance in multiple group confirmatory factor analysis, suggested by Schroeders, 2009 and Schroeders & Wilhelm, 2011, 2010). If the assumption of measurement invariance does not hold, it remains unclear whether this results from only some items showing a mode effect or from a general shift in difficulty. Manifest approaches, such as comparisons of means (e.g., Alexander, Bartlett, Truell, & Ouwenga, 2001; Bodmann & Robinson, 2004; Pommerich, 2004; Pomplun & Custer, 2005; Pomplun et al., 2002; Puhan et al., 2007; Schwarz et al., 2003; Wang, 2004) and cross-mode correlations of item parameters (Bennett et al., 2008; Pommerich, 2007), which are frequently described as ways of identifying differences between modes, are not sufficient to falsify hypotheses on equivalence without evidence of equal latent constructs, a fact which is often ignored.

The multiple-group IRT modelling approach proposed in this paper provides insights into item-specific mode effects by introducing a mode effect parameter. This parameter can apply either to all items or vary across items. When it varies across items, it is possible to find a selection of items that are invariant between modes. Such items can be used as anchor items in non-randomly equivalent groups. Furthermore, our approach helps to identify item properties that may increase the probability of mode effects.

In the next section, criteria for falsifying equivalence are presented, followed by the latent variable modeling approach that is illustrated for a within-subject and a between-subject design. This approach investigates mode effects regarding item parameters on both the test and item levels. Thereby, the criteria of equivalence related to item parameters, that is, (partial) measurement invariance, can be tested systematically.

Criteria of equivalence

A mode effect study attempts to provide empirical evidence justifying cross-mode comparisons. Criteria reflecting cross-mode equivalence should be specified on the basis of the intended use of test scores and related inferences. Thus, a mode effect study can be understood as part of the validation of the test score interpretation (cf. AERA, APA, & NCME, 2014). Intended comparisons differ for large-scale assessments, individual assessments and high-stakes tests. In large-scale assessments, typical uses of scores primarily include the comparison of means and correlations at the level of populations or subpopulations (Oliveri & von Davier, 2011; Rutkowski et al., 2013), whereas assessments on the individual level, including high-stakes testing, compare individual scores and often have relevant consequences for the test-taker.

The first criterion of equivalence is construct equivalence, that is, whether the construct measured by the test is the same in both modes. Despite its importance, this criterion has

received little consideration up to this point. The second and rather minor criterion of equivalence in some contexts is equal test reliability in both modes, which ensures that test score comparisons are not affected by differences in measurement accuracy. The equality of item parameters can be considered as a third criterion, and its investigation depends on the measurement model, which is described in the next section.

Measurement model and construct equivalence. A first step of a mode effect analysis is to determine an appropriate measurement model that fits simultaneously for both modes. In a within-subject design, a multi-dimensional IRT model has to be tested, while in the case of a between-subject design, a multiple-group IRT model needs to be tested and compared with respect to information criterion and used to investigate item fit. A measurement model that fits data from both modes simultaneously implies that the mode effects can be described as differences in the set of item parameters included in this measurement model (e.g., item difficulty and discrimination). Thus, the determination of a measurement model for both modes is prerequisite to absorb mode effects on IRT item parameters using latent variable modelling.

The question of construct equivalence – that is, whether the test captures the same latent variable in both modes is the first equivalence criterion. Construct equivalence is related to the step of determining a common measurement model: A latent variable model that includes responses from both modes enables testing construct equivalence by estimating latent correlations. When switching to another mode, construct-irrelevant individual differences or even another construct may be tapped, meaning that construct equivalence has to be tested when comparing results assessed in different modes (AERA, APA, & NCME, 2014; Huff & Sireci, 2001; ITC, 2005; Parshall et al., 2002; Penfield & Camilli, 2007; Puhan et al., 2007; Russell, Goldberg, & O'Connor, 2003). The technical implementation of hypotheses regarding construct equivalence depends on the specific study design (see section of design considerations). In large-scale assessments, where interest centers on comparisons of means and correlations with regard to a certain construct at the level of (sub-)populations, construct equivalence of the test versions is critically important.

Since most data in educational measurement are categorical, we restrict this step of determining an appropriate measurement model to IRT-based approaches, although categorical data can also be modelled within the framework of structural equation modeling (e.g., for mode effect analysis with CFA, Schroeders & Wilhelm, 2011). The IRT model to be chosen depends on whether item scoring is dichotomous or polytomous. The selected model should be as liberal in terms of item parameters so as to describe data from both modes appropriately. If the data from both modes do not fit to the same IRT model, a combined and more complex and thus liberal IRT model (e.g., integrating those items conforming to the Rasch model (Rasch, 1960) and those with deviating discriminations) can be used, as was, for instance, done in PISA (OECD, 2014b). In order to find an appropriate measurement model, another and frequently described possibility is excluding items that lead to worse model adjustment due to item misfit. Although this is one way of improving model fit, it limits the results of a mode effect study because the mode difference is only investigated for a subset of items (it could be the case that items with a large mode effect were excluded in pre-analysis).

To test for construct equivalence, latent or manifest cross-mode correlations can be used in a within-subject design (e.g., Mead & Drasgow, 1993). Manifest correlations are attenuated and can underestimate the true linear relationship if the constructs are measured with error. Therefore, latent correlations are expected to be not significantly different from 1 (r_1, Figure 2) if the order of the test parts is balanced, whereas manifest correlations are expected to be as high as predicted by the test's reliability in order to support the hypothesis of construct equivalence. An additional approach besides cross-mode correlations is to investigate the relation to external criteria. This is the method of choice in the case of a between-subject design, where cross-mode correlations between the latent variables are not possible due to different examinees responding to items in only one mode. Figures 2 and 3 show the correlations with an external criterion in a within-subject and between-subject design, respectively. The criterion of construct equivalence requires equal correlations (latent or manifest) of PBA with an external criterion (r_2, either latent or manifest) and CBA with the same external criterion (r_3, either latent or manifest). When estimating latent correlations, item parameters are freely estimated between modes because measurement invariance is not a precondition for construct equivalence. Modeling data with one IRT model and testing construct equivalence is necessary condition for subsequent steps such as concurrent calibration, which aligns item parameters along a common metric and thus allows scores to be used interchangeably.

Reliability. That both tests have same level of reliability is another criterion that needs to be ensured if results from different modes are to be compared (AERA, APA, & NCME, 2014; Holland & Dorans, 2006; ICT, 2005; Kolen & Brennan, 2004). This is more important in individual assessments, and can be considered optional in large-scale studies, as larger sample sizes may compensate for a decrease in reliability (see Adams, 2005, for reliability as a measurement design effect). In addition to individual assessments, equal reliabilities also become important when equating or linking between modes (Dorans, Moses, & Eignor, 2010) and when modes are changed in longitudinal studies (Buerger et al., 2015). In the context of IRT modeling, reliability is a function of item difficulty and item discrimination parameters, which represent another criterion for evaluating mode differences.

Item parameters. A third criterion according to which mode differences can be analyzed is item parameters, i.e. difficulty and discrimination (depending on the measurement model). Mode-related differences reflected in item parameters can be classified as a) homogeneous effects on the test level that affect all items in a similar manner or b) heterogeneous effects on the item level. Mode effects that vary across items might be systematic and depend on specific item properties, for example (Green, Bock, Humphreys, Linn, & Reckase, 1984). Analyzing mode effects on item parameters on the test and item level is described in the next section.

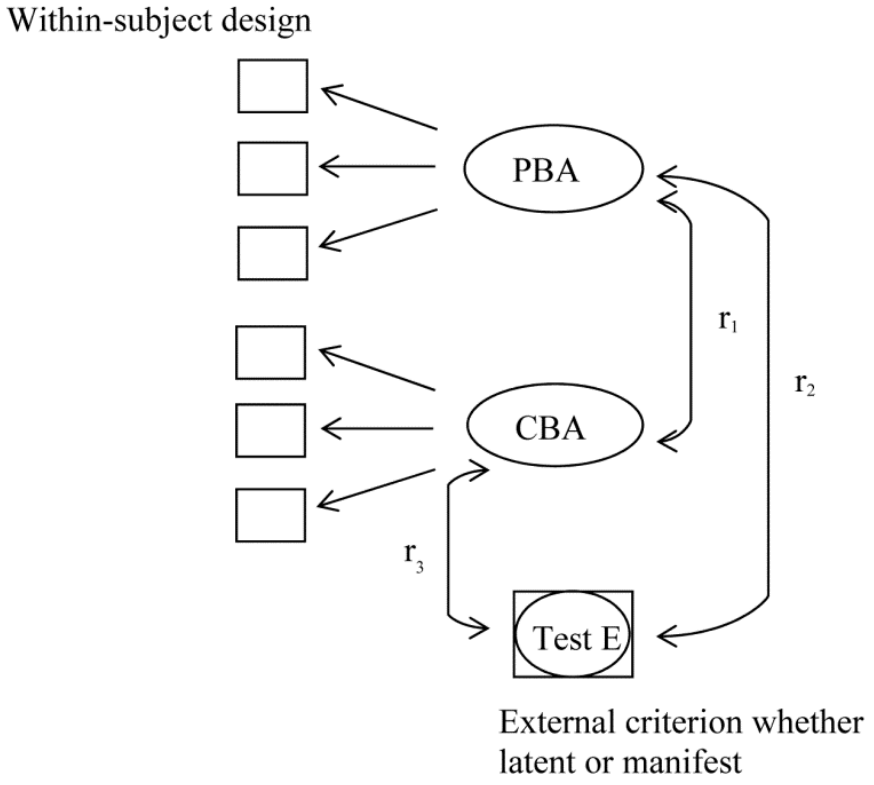


Figure 2:
Testing for construct equivalence in a within-subject design; r_2 is the correlation of PBA and the external criterion (Test E), r_3 is the correlation of CBA and the same external criterion and r_1 is the cross-mode-correlation

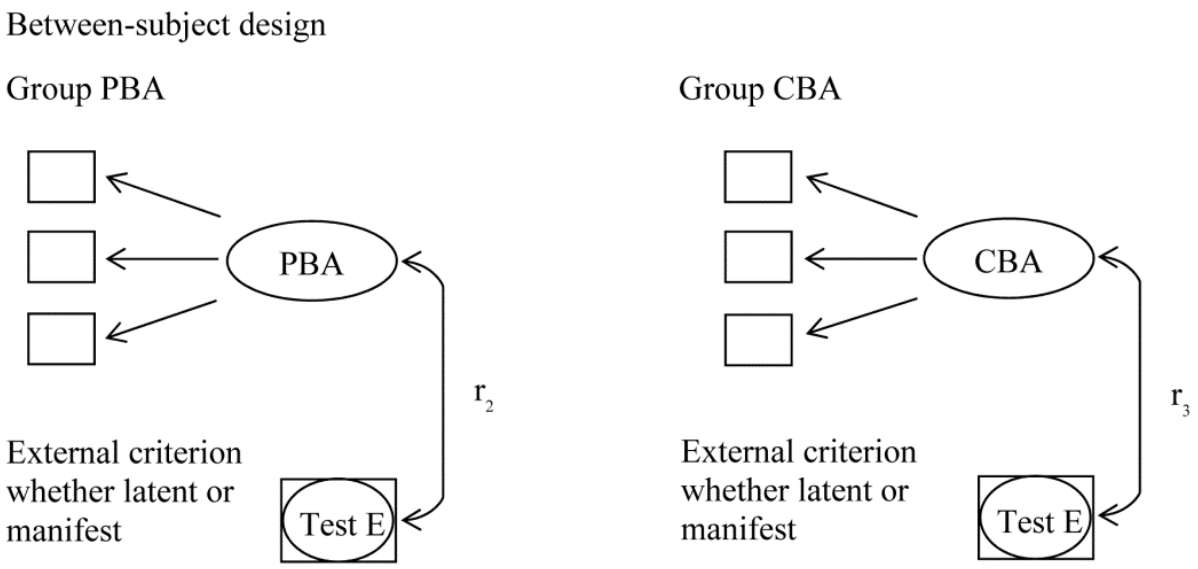


Figure 3:
Testing for construct equivalence in a between-subject design; r_2 is the correlation of PBA and the external criterion (Test E) and r_3 is the correlation of CBA and the same external criterion

The modeling approach

After investigating the common measurement model as a precondition for comparing item parameters between modes, measurement invariance has to be tested, that is, whether the relationship between the observed responses and the underlying latent variable is identical in both modes.

(Partial) Measurement invariance. Measurement invariance (i.e., equality of item parameters between groups) is typically addressed in studies in which groups are not randomly equivalent because an experimental assignment of persons to groups (defined by person-level variables) is not possible. For instance, when a PISA test is translated into multiple languages and differential item functioning between those test versions is investigated, persons cannot be arbitrarily assigned to language versions regardless of the person-level variable mother tongue. In mode effect studies, however, assignment to mode groups proceed randomly using the experimental design described above. This offers the advantage that measurement invariance (i.e., item parameters are invariant across both modes) is not needed to align items from both modes to one common metric, because performance differences cannot be the result of group differences but rather can be clearly attributed to differences in the administration mode.

The multiple-group IRT model approach allows for the investigation of partial measurement invariance. Specifically, partial measurement invariance is needed when the assumption of random equivalent groups does not hold (e.g., when persons are allowed to choose the mode themselves). In such cases, items that turn out to be invariant between modes can serve as anchor items. Kolen and Brennan (2004) describe requirements for common items to serve as an anchor. They have to represent the measured construct and must be representative of the test's specifications. That is, the more heterogeneous the test content is, the more invariant items are needed to capture variety in content. In general, the random equating error decreases with an increasing number of invariant items. In order to avoid a distortion in equating due to position effects, the position of the anchor items should be the same in both test versions.

Model. The common measurement model that fits the CBA and PBA data determines the item parameters and thereby where differences between administration modes can be observed. If the Rasch model applies, discrimination is assumed to be equal for all items within groups, so that only item difficulty is investigated in equivalence analysis. In a 2 PL model, item discrimination also needs to be examined for equivalence, while in a 3 PL model the guessing parameter needs additional consideration (Birnbaum, 1968). In this paper, we illustrate the investigation of partial measurement invariance for tests that fit the Rasch model, meaning that a potential mode effect only affects item difficulties. However, the described modelling approach can be easily generalized to more complex IRT models.

In the Rasch model, the probability P of a correct response of person *i* on item *j* in mode (group) *m* is defined as:

$$logit\left[\mathrm{P}\left(Y_{ijm}=1\right)\right]=\theta_i-\beta_{jm} \tag{1}$$

In the case of a between-subject design, the mode *m* is the mode group to which test-takers are randomly assigned. In a within-subject design, the mode *m* stands for the order-balanced test parts in PBA and CBA.

Under PBA (mode = 0), item *j* has the difficulty of $\beta_{j,PBA}$:

$$logit\left[\mathrm{P}(Y_{ijm}=1|G=0)\right]=\theta_i-\beta_{j,PBA} \tag{2}$$

If there is a mode effect at the item level, for item *j* the difficulty is changed to the amount of a new introduced mode parameter ME_j under CBA (mode = 1):

$$logit\left[\mathrm{P}(Y_{ijm}=1|G=1\right]=\theta_i-(\beta_{j,PBA}+ME_j) \tag{3}$$

If there is a systematic mode effect across all CBA items (i.e., mode effect at the test level), all items show the same shift in difficulty, that is, ME_j is equal for all items *j*:

$$logit\left[\mathrm{P}(Y_{ijm}=1|G=1)\right]=\theta_i-(\beta_{j,PBA}+ME) \tag{4}$$

With this multiple-group model approach introduced, invariant items are identified by testing differences in item parameters between modes. To do this, software is required that allows for the modeling of multiple groups and the estimation of standard errors of the difference between item parameters from different groups. This can be done, for example, in Mplus (Muthén & Muthén, 1998-2015) using the MLR estimator and mixture type of analysis, and by introducing a new parameter (ME_j) representing the difference in the item parameters for all item pairs from PBA and CBA. To test mode effects, constraints are imposed on this new parameter ME_j. Basically, if the constraints worsen the model fit, meaning that equality assumptions do not hold, a mode effect is indicated.

Testing constraints. To test for measurement invariance using our modelling approach, three hypotheses on the test and item levels are inspected. *Hypothesis A* is that no mode effect exists at the test level. In *Hypothesis B,* the mode effect is expected to be equal for all items, whereas in *Hypothesis C* it varies across items. *Hypothesis A* is tested by the constraint $\sum ME_j=0$, which means that across all items *j= 1...J* there is no mode effect. To test this hypothesis, the Wald test statistic (Wald, 1943) can be used. A significant Wald test statistic for *Hypothesis A* means that there is a mode effect on item difficulties. Here, it is important to note that the sum of all item differences might also be zero and the Wald test insignificant if there are some mode effects in opposite directions that cancel each other out. Finding a significant mode effect in this step means there has been a shift in average item difficulty from one mode to another. Note that for the identification of the mode specific parameter $ME_{j,}$ we either need the assumption of random equivalent groups, which allows us to fix the mean of the latent variable in both mode groups to zero, or anchor items must be defined to allow for the estimation of latent mean differences. In this illustration of the modelling approach, we assume random equivalent groups. Furthermore, as a consequence of the assumption of the Rasch model

as the IRT model that fits the data in both modes, the variances of the latent variable can be constrained to be equal.

The next step on the test level is to analyze whether this mode effect is homogeneous over all items. Therefore, in *Hypothesis B,* all differences between item parameters are constrained to be equal across items, $ME = ME_j$. If *Hypothesis B* holds, all items on the computer get a CBA-specific item parameter by adding the *ME* component to the PBA-specific item parameter $\beta_{j,CBA} = \beta_{j,PBA} + ME$. In the case where *Hypothesis B* is rejected, the mode effect cannot be simplified to a general shift in item difficulties, and the model with item-specific mode effects fits the data better. Accordingly, *Hypothesis C* tests mode effects on the item level for each item *j*: $ME_j = 0$. For each item, the mode effect is represented by the difference in PBA and CBA item parameters that is tested to be different from zero. Given the estimated standard error for the difference between the item difficulties in both modes, a t-test can be conducted for each item *j*, testing whether or not ME_j is different from zero. This test can be used to identify single items showing a mode effect, thus also identifying anchor items. If there is a mode effect for only some items in the test, those items get a specific item parameter for CBA, $\beta_{j,CBA} = \beta_{j,PBA} + ME_j$. Partial measurement invariance can be assumed for those items for which *Hypothesis C* holds and that are thus not affected by the mode change.

Furthermore, the multiple-group IRT model approach allows for an investigation of the mode effect at the item level by relating it to item properties (as previously described in the section on sources of mode effects). To do this, a new parameter is created, representing whether a given item exhibits such a property. For instance, all computerized items with the need for scrolling could be assumed to be more difficult under CBA than PBA. Thus, the mode effect at the item level can be explained by a set of item properties that are assumed to induce a mode effect: $ME_j = \gamma_1 \bullet \mathrm{x}_{j1} + \gamma_2 \bullet \mathrm{x}_{j2} + \ldots + \gamma_k \bullet \mathrm{x}_{jk}$. For each item property *k*, x_{jk} indicates whether this property is given for item *j* with $\mathrm{x}_{jk} = 1$ and $\mathrm{x}_{jk} = 0$ otherwise. If this decomposition can be identified, the weight γ_k indicates how strong the property contributes to the mode effect. For instance, if the property "scrolling" is given for an item, and it shows a significant effect on the logit scale, this effect can be translated into a change in the probability of completing the item successfully.

Discussion

Since a change from paper-based to computer-based assessment has taken place in many large-scale assessments, including, for instance, PISA, PIAAC, and NEPS, mode effect studies are highly relevant. In mode effect studies between 2000 and today, both the way tests are presented on the computer and the methods used to check equivalence vary considerably. Some researchers made decisions about equivalence solely by comparing mean scores for mode groups with no regard to items, whereas others did an extensive

investigation of item parameters, the order of test versions and characteristics of test-takers or subgroups. In addition, differences in sample size and thus the power of statistical tests have let some effects of the administration mode be detected, while others probably could not be discovered. Thus, empirical evidence about the equivalence of modes cannot be compared easily. For this reason and to propose a standard tool, respectively, we presented a multiple-group IRT model approach for investigating differences between paper-and-pencil and computer tests in item parameters. Moreover, we discussed one major equivalence criteria, that is, the issue of construct equivalence.

Defining the measurement model and checking for construct equivalence has seldom been examined in previous studies, although it is critical to the interpretation of data from different administration modes. The step of defining a measurement model determines the parameters in which mode differences may be observed. Regarding differences in item parameters between modes, we revert to the terminology of measurement invariance testing (where groups usually are non-randomly equivalent). Transferring this approach to a mode effect analysis assuming randomly equivalent groups generated via experimental designs offers an opportunity to investigate partial measurement invariance. If groups are non-randomly equivalent for some reason, invariant (anchor) items are needed to put scores on a common metric and use scores interchangeably.

Examining each item for mode effects allows the model to be expanded easily to investigate whether the mode effect depends on item properties (e.g., response format). As part of this process, item properties that increase the difficulty of an item on the computer can be identified. Knowing about those properties provides an opportunity to construct test items that will be less prone to mode effects. To better understand and prevent differences between modes, further research on how item properties are related to mode effects is required (Buerger et al., 2015). Characteristics of subgroups or countries disadvantaged by a particular administration mode should also be considered in future mode effect analyses of large-scale assessments.

If significant mode effects are found for specific items, the question arises of how big the effect size is and whether this effect is of practical relevance. The increase (or decrease) of probability of success for test-takers with the same ability but taking the test in different modes may help to illustrate the relevance of an effect. However, the literature on DIF could be consulted to evaluate effect sizes more clearly, and techniques proposed there could be adapted to the analysis of mode effects (see Magis, Béland, Tuerlinckx, & de Boeck, 2010, for a detailed overview of appropriate methods as well as Zieky, 1993 for a classification scheme).

Some research has shown that computer-based tests have a faster completion time (e.g., Alexander et al., 2001; Bodmann & Robinson, 2004), which means that more items can be presented to test-takers on the computer than on paper. This should be considered if reliability differs between modes because the additional items might compensate for reliability differences by increasing the reliability of the computer test. Differences in time intensity can be regarded as the result of a mode effect, or as a variable mediating an effect on other psychometric properties such as difficulty. Here, further analysis in-

vestigating test-taking time for computer-based assessments compared to paper-based tests is required.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31,* 162-172.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington: AERA, APA, NCME.

Alexander, M. W., Bartlett, J. E., Truell, A. D., & Ouwenga, K. (2001). Testing in a Computer Technology Course: An Investigation of Equivalency in Performance Between Online and Paper and Pencil Methods. *Journal of Career and Technical Education*, *18* (1). Retrieved from http://scholar.lib.vt.edu/ejournals/JCTE/v18n1/alexander.html

American Psychological Association, Committee on Professional Standards (COPS), & Committee on Psychological Tests and Assessments (CPTA). (1986). *Guidelines for computer-based tests and interpretations*. Washington: American Psychological Association, Inc.

Association of Test Publishers (ATP). (2000). *Guidelines for computer-based testing*. Washington DC: Association of Test Publishers.

Bennett, R.E. (2002). *Using electronic assessment to measure student performance*. (Issue Brief). Washington, DC: NGA Center for Best Practices. Retrieved from http://www.nasbe.org/Standard/10_Summer2002/bennett.pdf

Bennett, R. E. (2003). Online Assessment and the Comparability of Score Meaning. *Research Memorandum 3* (5), 1–19. Retrieved from http://www.ets.org/Media/Research/pdf/RM-03-05-Bennett.pdf

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. *The Journal of Technology, Learning, and Assessment, 6* (9).

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examiner's ability. In F. M. Lord & M. R. Novick (Ed.) *Statistical theories of mental test scores,*17–20.

Blossfeld, H.-P., Roßbach, H.-G, & von Maurice, J. (Eds.) (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). [Special Issue] *Zeitschrift für Erziehungswissenschaft, 14*.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research, 31* (1), 51–60.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (ETS-RR-01-23). Princeton, NJ: Educational Testing Service.

Buerger, S., Kroehne, U., & Goldhammer, F. (2015, July). *Investigating Mode Effects in Reading Assessments.* Paper presented at the 6th Conference of the European Survey Research Association, Reykjavik.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology, 33* (5), 593–602.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and Practices of Test Score Equating* (ETS Research Rep. No. RR-10-29). Princeton, NJ: ETS.

Eid, M. (2006). Methodological approaches for analyzing multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of psychological measurement: A multimethod perspective* (pp. 223–230). Washington, DC: American Psychological Association.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical Guidelines for Assessing Computerized Adaptive Tests. *Journal of Educational Measurement, 21* (4), 347–360. Retrieved from http://www.jstor.org/stable/1434586

Heerwegh, D., & Loosveldt, G. (2002). An Evaluation of the Effect of Response Formats on Data Quality in Web Surveys. *Social Science Computer Review, 20* (4), 471–484.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the Effect of Computer-Based Passage Presentation of Reading Test Performance. *The Journal of Technology, Learning, and Assessment, 3* (4).

Holland, P. W., & Dorans, N. J. (2006). Linking and Equating. In R. L. Brennan (Ed.), *Educational measurement (4th ed.)* (pp. 187–220). Westport, CT: Praeger.

Hox, J.J., de Leeuw, E. D., & Zijlmans, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology, 6*:87. doi: 10.3389/fpsyg.2015.00087

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practices, 20* (3), 16–25.

International Test Commission (ITC). (2005). *International Guidelines on Computer-Based and Internet Delivered Testing.* Retrieved from https://www.intestcom.org/files/guideline_computer_based_testing.pdf

Kim, D.-H., & Huynh, H. (2008). Computer-Based and Paper-and-Pencil Administration Mode Effects on a Statewide End-of-Course English Test. *Educational and Psychological Measurement, 68* (4), 554–570.

Kingston, N. M. (2009). Comparability of Computer- and Paper-Administered Multiple-Choice Tests for K–12 Populations: A Synthesis. *Applied Measurement in Education, 22,* 22–37.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.)*. New York: Springer-Verlag.

Kroehne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft, 14* (2), 169–186.

Magis, D., Béland, S., Tuerlinckx, F, & de Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42* (3), 847-862. doi:10.3758/BRM.42.3.847

Mazzeo, J., & Harvey, A. L. (1988). *The Equivalence of Scores from Automated and Conventional Educational and Psychological Tests. A Review of the Literature* (College Board Report 88-8). New York: College Entrance Examination Board.

Mazzeo, J., & von Davier, M. (2008). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. *Education Working Papers EDU/PISA/GB (2008)*, 28, 23–24.

Mead, A. D., & Drasgow, F. (1993). Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. *Psychological Bulletin, 114* (3), 449-458.

Muthén, L.K., & Muthén, B.O. (1998-2015). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

OECD (2013). The Survey of Adult Skills: Reader's Companion, OECD Publishing. http://dx.doi.org/10.1787/9789264204027-en

OECD (2014a). PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014), PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201118-en

OECD (2014b). Pisa 2015 Field Trial Goals, Assessment Design, and Analysis Plan for Cognitive Assessment, PISA, OECD Publishing.

OECD. (2015). Adults, Computers and Problem Solving: What's the Problem? OECD, Publishing. http://dx.doi.org/10.1787/9789264236844-en

Oliveri, M. E., & von Davier, M. (2011). Investigating of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53* (3), 315-333.

Osterlind, S. J., & Everson, H. T. (2009). *Differential Item Functioning*. Thousand Oaks: SAGE Publications, Inc.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2010). Innovative Items for Computerized Testing. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Elements of Adaptive Testing, Statistics for Social and Behavorial Sciences* (pp. 215–230).

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Pearson Educational Measurement. (2005). *Recent Trends in Comparability Studies*. PEM Research Report 05-05. Austin, TX: Author

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26. Psychometrics,* (pp.125–167). New York, NY: Elsevier.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large Scale State Assessment Program. *The Journal of Technology, Learning, and Assessment, 3* (6).

Pommerich, M, & Burden, T. (2000, April). *From Simulation to Application: Examinees React to Computerized Testing.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

Pommerich, M. (2004). Developing Computerized Versions of Paper-and-Pencil Tests: Mode Effects for Passage-Based Tests. *The Journal of Technology, Learning, and Assessment, 2* (6), 3-44.

Pommerich, M. (2007). The effect of using item parameters calibrated from paper administrations in computer adaptive test administrations. *The Journal of Technology, Learning, and Assessment, 5* (7). Retrieved from http://files.eric.ed.gov/fulltext/EJ838609.pdf

Pomplun, M., & Custer, M. (2005). The Score Comparability of Computerized and Paper-and-Pencil Formats for K-3 Reading Tests. Journal of Educational Computing Research, 32 (2), 153-166.

Pomplun, M., Frey, S., & Becker, D. F. (2002). The Score Equivalence of Paper-and-Pencil and Computerized Versions of a Speeded Test of Reading Comprehension. *Educational and Psychological Measurement, 62,* 337–354.

Puhan, G., Boughton, K., & Kim, S. (2007). Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. *The Journal of Technology, Learning, and Assessment, 6* (3). Retrieved from http://www.jtla.org

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87* (3), 517–529.

Rasch, G. (1960). *Probabilistic Models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114* (3), 552–566.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2013). *Handbook of International Large-Scale Assessment: Backround, Technical Issues, and Methods of Data Analysis*. CRC Press, Boca Raton.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-Based Testing and Validity: A Look Back and Into the Future. *Assessment in Education, 10* (3), 279-293.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of Reading and Listening Comprehension Across Test Media. *Educational and Psychological Measurement, 71* (5), 849–869. Retrieved from http://epm.sagepub.com/content/71/5/849

Schroeders, U., & Wilhelm, O. (2010). Testing Reasoning Ability with Handheld Computers, Notebooks, and Paper and Pencil. *European Journal of Psychological Assessment, 26* (4), 284–292.

Schroeders, U. (2009). Testing for equivalence of test data across media. In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment. Lessons learned from the PISA 2006 computer-based assessment of science (CBAS) and implications for large scale testing* (pp. 164–170).

Schwarz, R. D., Rich, C., & Podrabsky, T. (2003, April). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests*. Paper presented at Annual Meeting of the National Council on Measurement in Education, Chicago.

Sireci, S.G., & Zenisky, A. L. (2006). Innovative Item Formats in Computer-Based Testing: In Pursuit of Improved Construct Representation. In Downing, S. M. & Haladyna, T. M. (Eds.), *Handbook of test development* (pp. 329–347).

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models. *Journal of Educational and Behavioral Statistics, 30* (4), 443–464.

Vispoel, W. P. (2000). Reviewing and Changing Answers on Computerized Fixed-Item Vocabulary Tests. *Educational and Psychological Measurement, 60* (3), 371–384.

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is large. *Transactions of the American Mathematical Society*, *54* (3), 426-482.

Wang, S. (2004). *Online or Paper: Does Delivery Affect Results? Administration Mode Comparability Study for Stanford Diagnostic Reading and Mathematics Tests*, San Antonio, Texas: Harcourt.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K 12 Reading Assessments: A Meta-Analysis of Testing Mode Effects. *Educational and Psychological Measurement, 68* (1).

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria, and an example. *Journal of Educational Measurement, 38,* 19-49.

Yamamoto, K. (2012). *Outgrowing the Mode Effect Study of Paper and Computer Based Testing.* Retrieved from http://www.umdcipe.org/conferences/EducationEvaluationItaly/COMPLETE_PAPERS/Yamamoto/YAMAMOTO.pdf

Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

Anhang B

Beitrag II: Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (*submitted, Computers in Human Behavior*). Construct Equivalence of PISA Reading Comprehension Measured with Paper-based and Computer-based Assessment.

Construct Equivalence of PISA Reading Comprehension Measured with Paper-based and Computer-based Assessment

Ulf Kroehne[a], Sarah Buerger[a], Carolin Hahnel[ab] & Frank Goldhammer[ab]

[a] German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany

[b] Centre for International Student Assessment (ZIB), Germany

Author Note

Ulf Kroehne, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany; Sarah Buerger, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany; Carolin Hahnel, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany; Frank Goldhammer, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany.

Correspondence concerning this article should be addressed to:

Ulf Kroehne, German Institute for International Educational Research (DIPF),

Solmsstraße 73-75, 60486 Frankfurt am Main, Germany.

Phone: +49 69 24708 728

E-mail: kroehne@dipf.de

Abstract

For several years, reading comprehension has been measured with paper-based assessment (PBA) in the Programme for International Student Assessment (PISA). In the latest cycle in 2015, computer based assessment (CBA) has been introduced, raising the question of whether test equivalence criteria required for a valid interpretation of results are fulfilled. As an extension of the PISA 2012 main study in Germany a random subsample was assessed with two intact PISA reading clusters, either computerized or paper-based using a random-group design with an additional within-subject variation. Results indicate that the hypothesis of construct equivalence can be retained. That is, the latent cross-mode correlation of PISA reading comprehension was not statistically significant different from the expected correlation of the two clusters. Significant mode effects were observed on item difficulties for a small number of items only. Using a model developed for method effects in multitrait-multimethod data inter-individual differences between CBA and PBA were investigated. For this, a latent difference variable was estimated using multiple-group structural equation models. Inter-individual differences in the mode effect were negatively correlated with reading comprehension, but neither predicted by basic computer skills nor gender. Further differences between modes were found with respect to the number of missing values.

Keywords: Mode Effects; Construct Equivalence; Computer-based assessment; Reading Comprehension; Programme for International Student Assessment

1. Introduction

The Programme for International Student Assessment (PISA) shifted to computer-based testing in all domains in 2015 (OECD, 2016). Subsequent to the publication of the results some authors found indicators that the cross-sectional comparisons as well as the comparability of estimates to previous assessments (i.e., trend estimates) were probably influenced by effects of the change from paper-pencil to computer tests ("mode effects"; e.g., Klieme 2016, Komatsu & Rappleye, 2017, Robitzsch et al. 2017, and also Jerrim, 2016), questioning the effective handling of potential mode effects in the procedures applied to the PISA 2015 cognitive measures.

Mode effect research and the investigation of test equivalence of achievement and competence tests has a 50 years long tradition in different areas of psychological diagnostics and educational assessment (early publications dating back to the late 1960ies, e.g., Vinsonhaler, Molineaux, & Rodgers, l968). Although international large-scale assessments (ILSAs) like the Programme for the International Assessment of Adult Competencies (PIAAC) have already introduced computer-based assessment (CBA, see OECD, 2015, and Yamamoto, 2012), the topic of mode effects is still relevant for educational research: Other ILSAs, such as the Trends in International Mathematics and Science Study (TIMSS) as well as national large-scale assessments such as the National Assessment of Educational Progress (NAEP) in the U.S. and the National Educational Panel Study (NEPS) in Germany are still in the process of replacing paper-based assessment (PBA) with CBA.

This paper deals with the comparability of PBA and CBA regarding identical PISA items for the domain of reading, the major domain in the upcoming PISA 2018 assessment. It adds to the current literature by reviving the distinction between the two problems that are both labeled as *mode effects*: effects on the measurement model versus effects on the assessed construct (Mead & Drasgow, 1993, p. 449). We present empirical results from the cohort of 15-year-old students assessed in PISA 2012 in Germany, and provide methodological insights

into the investigation of construct equivalence using data from a within-subject design. Mode effects are investigated at item and test level using multi-group structural equation models. Furthermore, it is examined whether inter-individual differences in the mode effect are correlated with reading and covariates such as basic computer skills.

In the next section previous research about mode effects for reading assessments and the treatment of mode effects in PISA 2015 are briefly reviewed. This will be followed by a section about the theoretical background for investigating the comparability of PBA and CBA including the definition and identification of mode effects. In the subsequent section equivalence criteria for large-scale assessments and a specific psychometric model developed for multitrait-multimethod data will be described. The introductory part closes with the presentation of five hypotheses on the nature of mode effects in PISA reading comprehension that will be investigated in the empirical part of the paper.

2. Selected Previous Results on Mode Effects

Previous research on test equivalence often questioned the existence of mode effects. In their well-known meta-analysis Mead and Drasgow (1993) concluded, that "*there is no medium effect for carefully constructed power tests"* (p. 457). As this seems to contradict, for instance, the above-mentioned doubts about mode effects in PISA 2015 (Robitzsch et al., 2017; Komatsu & Rappleye, 2017), Table 1 lists key challenges in research on mode effects that might limit the value of previous studies for predicting potential mode effects for the PISA assessment.

Requirement	Reason
Identical Sample Composition	• Mode effects should be investigated in samples representative with respect to the target population
Valid Cohort	• Mode effects might change over time but findings must be (still) valid for the cohort of interest
Sufficient Large Sample Size	• Appropriate statistical power needed to falsify the (desired) null hypothesis of equivalence
Justified Equivalence Criteria	• Equivalence criteria depend on the intended use of the score • Effects on item parameters versus effects on the measured construct should be distinguished • Investigating invariance of item parameters requires fit of a common measurement model • Additional criteria required, e.g., o Number of missing values / type of missing values o Test-taking effort and test-taking engagement o Response times and test speededness
Appropriate Study Design	• Between-group designs require external criteria to assess construct equivalence • Within-group designs allow estimating cross-mode correlations by requiring subjects to complete selected test items in different modes
Detailed Treatment Description	• Detailed information about the specific implementation of the investigated computer-based implementation • Required to judge the comparability of properties of the investigated test administration

Table 1. Key methodological requirements for mode effect analysis.

An exhaustive review of previous research is beyond the scope of this paper. However, as the results of Robitzsch et al. (2017) give reasons to assume domain specific results we present a more detailed review of previous results on the comparability of reading assessments.

2.1 Previous Results on the Comparability of Reading Assessment

The nature of reading assessments (i.e., a reading text probably with multiple pages followed by several questions) makes reading assessments specifically prone to mode effects compared to mathematics assessments (e.g., Pommerich, 2004). Previous empirical studies showed heterogeneous results regarding the existence of mode effects (Noyes & Garland, 2008; Wang, Jiao, Young, Brooks, & Olson, 2008). Several studies discussed intratextual

navigation as a potential reason for mode effects in reading assessments (e.g., Mangen, Walgermo, & Brønnick, 2013; Pommerich, 2004). In particular, scrolling has been identified as increasing the difficulty of items on computer (e.g., Bridgeman, Lennon, & Jackenthal, 2003; Kim & Huynh, 2008; Poggio, Glasnapp, Yang, & Poggio, 2005; Wang et al., 2008).

Mode effects appear to be complex, and are likely to depend on the particular testing program (Kolen & Brennan, 2014). They even might not be generalized beyond different computer interfaces (Pommerich, 2004). Hence, even after many years of research, the empirical evidence regarding the existence and size of mode effects is still inconclusive (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) and thus, mode effect studies are of continuous relevance.

2.2 Treatment of Mode Effects in PISA 2015

OECD (2016, Annex A6) describes how responses from CBA and PBA were used to attempt comparability of trend estimates between PISA 2015 and previous cycles. The applied methodology involved at least three conceptually distinguishable steps: 1) the identification of items affected by the change of mode using data from an experimental design component in the PISA 2015 field trial, 2) the incorporation of the information, which items were found to be invariant in the field trial analysis of step one, into the analysis of the PISA 2015 main assessment and 3) additional adjustments incorporated in the scaling procedure based on item fit statistics in the concurrent calibration using data from the current and previous cycles (historical PISA database, which includes all prior assessment data across the 2000-2012 PISA cycles).

The PISA 2015 field trial was utilized in step one to investigate mode effects in identical items administered as PBA and CBA to random equivalent groups in a between-subject design implemented in 58 countries (OECD, 2016, Annex A6). Item parameter estimates obtained from separate calibrations of PBA and CBA reading data were compared and their

high correlations (reading domain 0.95 for difficulties and 0.90 for slopes) were interpreted implying that CBA and PBA measure the same construct. Moreover, a sequence of IRT models was estimated in mdltm (von Davier, 2005) that incorporates either no mode effect at all (Model 1), a homogenous shift in item difficulties (Model 2), item-specific shifts in item difficulties (Model 3) or an additional uncorrelated random variable that represents additional inter-individual differences weighted by item-specific mode-slope parameters (Model 4). To achieve parsimony, Model 3 was selected, although Model 4 was superior in terms of the considered information criteria (AIC, BIC and CAIC) for the reading domain (OECD, 2016, Annex A6). Until now it is unknown, how this additional variability introduced in Model 4 is related to the measured constructs and how the inter-individual differences in the mode effect might affect the interpretation of computer-based assessed competences. Basing on a theoretical foundation of mode effects and the deduction of equivalence criteria for large-scale assessments this paper aims at filling this gap for one particular domain.

3. Theoretical Foundation of Mode Effects

3.1 Properties of Test Administration

In order to define mode effects, we briefly provide the following theoretical background: Assessment can be conceptualized as items (i.e., item content such as stimulus and questions) administered to test takers in a particular way (i.e., in a particular mode, implemented in a specific manner). Properties of the test administration (PoTA) describe facets of the operational data collection (i.e., the specific manner of the implementation). Multiple PoTA are typically required to describe differences between assessments, for instance, between CBA and PBA.

3.2 Definition of Mode Effects

We define mode effects as differences in the measurement caused by unequal PoTA, while items (e.g., the stimulus including the reading text and questions) are identical. A specific assessment is always made up by a combination of different PoTA. Taking changes

in computer-based testing environments into account we generalize mode effects as effects of a specific combination of PoTA (including combinations that use the identical medium, e.g., two different CBA tests). The differences in the assessment caused by different PoTA is expected to result in an unknown mixture of effects such as, for instance, differences in item difficulties that partially might compensate or cancel-out each other. In line with Green, Humphreys, Linn and Reckase (1984) we distinguish three possibilities, how items can be affected by differences in PoTA:

- Type A: homogenous mode effect for all items;
- Type B: item by mode interaction (i.e., different mode effects for different items) or
- Type C: structural changes of the measured construct according to the different realizations of PoTA between modes.

Structural changes between a construct measured with one combination of PoTA (e.g., a new mode such as CBA) and another combination of PoTA (e.g., an old mode such as PBA or an outdated CBA implementation) might be associated with both relevant and irrelevant variance for the measurement of that particular construct (type C). Finally, specific PoTA known, for instance, from previous research, can be identified that are in particular relevant as potential sources for mode effects (see Kroehne & Martens, 2011, for details).

Properties of Test Administration (PoTA) that differ between CBA and PBA	**Mode (A)**	**Item by Mode (B)**	**Construct (C) / Inter-Individual Differences**
• Medium for the stimuli (affects reading behavior, probably reading speed and depth of processing)	all items		computer usage; reading competence; gender
• Time limits (strictness of holding time constraints, remaining effective time for answering questions after reading the test stimulus)	all items		individual speed level; test-taking strategies
• Technical process of transmitting the response (e.g., typing versus writing, selecting vs. crossing, writing a letter vs. choosing an entry from a combo box)		response format	computer usage and literacy; typing skills; handwriting usage
• Possibility to remove, delete, change or update a previously already given answer		response format	computer usage and literacy
• Distance between questions and text (same page vs. different screens for affected items)		position within units	short-term memory; computer usage and literacy
• Possibility to navigate within units (i.e., answering questions of a unit in an arbitrary self-selected order)	all items	position within units	test-taking strategies; computer usage and literacy
• Restricted navigation between units (within parts such as booklets, i.e., possibility to work on units in a self-selected order)	all items	position within units	test-taking strategies
• Possibility to put items aside for answering later (item review)	all items	item difficulty	test-taking strategies
• Missing value indicator notifying the test taker when items are not yet answered completely	all items	item-specific missing propensity	test-taking motivation; missing propensity
• Feedback about the number of answered items (compared to the number of not yet answered items) and about the available time	all items	position within booklet	test-taking motivation; missing propensity

Table 2: Selected properties of test administration that differ between PBA and CBA and a theoretical classification with respect to their possible effect on mode effects regarding the scale (A and B) and the measured construct (C)

As summarized in Table 2 it is possible to hypothesize whether differences in PoTA are likely to affect all items at the test-level (e.g., contributing to an overall shift of item difficulties; corresponding to type A) or items differently (i.e., item by mode interactions;

corresponding to type B) or which additional person-level covariate might reflect additional construct-irrelevant variance (i.e., potential change in the measured construct due to this PoTA; corresponding to type C). Note that neither Table 2 is complete nor that there is a known complete list of *all* properties that are necessary to describe differences between PBA and CBA.

3.3 Identification of Mode Effects

Mode effects defined as differences in the measurement caused by different composites of PoTA (for PBA vs. CBA) can be identified, for instance, using random assignment of test takers to modes. However, without experimental variation of single realization of PoTA it is impossible to disentangle the contribution of the different properties to the resulting mode effect. Even though the empirical phenomenon of mode effects can be analyzed using techniques developed for the analysis of differential item functioning (DIF), we emphasize an important conceptual difference: Mode effects can be identified using random assignment of test takers to administrations with different PoTA (e.g., mode groups PBA or CBA) to create independence of the mode with any measured or unmeasured variable (including the measured latent construct), while typical variables used to distinguish groups in DIF analysis (e.g. gender) cannot be assigned randomly. Consequently, the assumption of *pure* (link-) items (or anchor items) that behave similar between both modes is not necessary to identify mode effects.

Mode effect analysis should also be conceptually differentiated from *measurement invariance*. Measurement invariance, weak measurement invariance and factorial invariance are defined (e.g., Meredith, 1993) with respect to a *selection* on a variable that is by assumption not independent of the measured latent construct. Hence, referring to the terminology of degrees of measurement invariance (e.g., Schroeders & Wilhelm, 2011) can be used to communicate the relationship between measures using different PoTA, but it should

not mask the important difference: the possibility to investigate mode effects with experimental designs.

4. Equivalence Criteria

To apply the definition of mode effects provided above, it is necessary to specify appropriate equivalence criteria. Such criteria used for a specific mode effect investigation has to be derived from the intended use of scores of a particular assessment. Hence, specific equivalence criteria can be derived to falsify mode effects in ILSAs, such as for the PISA reading assessment. The comparability of population and sub-population estimates at the construct level requires at least *construct equivalence* as prerequisite for psychometric treatment of mode effects (e.g., Lottridge, Nicewander, Mitzel, & Metrics, 2010; Dorans & Cook, 2016). Hence, the absence of differences regarding construct irrelevant and construct relevant variance (McDonald, 2002; Mead & Drasgow, 1993) can be seen as a central equivalence criterion required for maintaining comparable score interpretations in PISA reading assessment, when results are interpreted at the level of (sub-) population estimates. Construct equivalence is also the prerequisite to meaningfully adjust individual level estimates for mode effects either using one common measurement model or using a specific linking or equating technique (e.g., Kolen & Brennan, 2014). The data necessary for such an adjustment could be gathered, either in a field trail or in the main assessment itself as soon as the mode effect is identified with a particular experimental design (random assignment). An example, how to apply linking based on the field trial data of PISA 2015 can be found in Robitzsch et al. (2017) following the idea of a bridge study (Mazzeo & von Davier, 2008), i.e., linking from previous cycles to PISA 2015 with the help of the randomized field trial.

As discussed in Buerger, Kroehne & Goldhammer (2016) the investigation of construct equivalence requires either a within-group design or a between-group design with additional criterion variables. To the best of our knowledge construct-equivalence has yet not been explicitly investigated for the PISA reading assessment in paper- and computer-based mode.[1]

4.1 Construct-Equivalence and Cross Mode Correlations

Investigating construct equivalence in a within-subject design requires a booklet design that allows estimating the cross mode correlation (i.e., the latent correlation of the identical construct between assessments differing in PoTA). For that purpose, the instrument can be split into different parts (e.g., two non-overlapping forms "A" and "B") and the booklet design must ensure that at least a sub-group of test takers answers one part of the instrument in one mode and another part in the other mode. To avoid administering identical items twice, no test taker answers identical items in both modes. To avoid confounding of position and mode, the booklet design should be balanced. The latent correlation between CBA and PBA (see ψ_{21} in panel "Latent Cross-Mode Correlation Model" in Figure 1) provides a criterion to judge the equivalence of the construct measured in different modes (Mead & Drasgow, 1993).

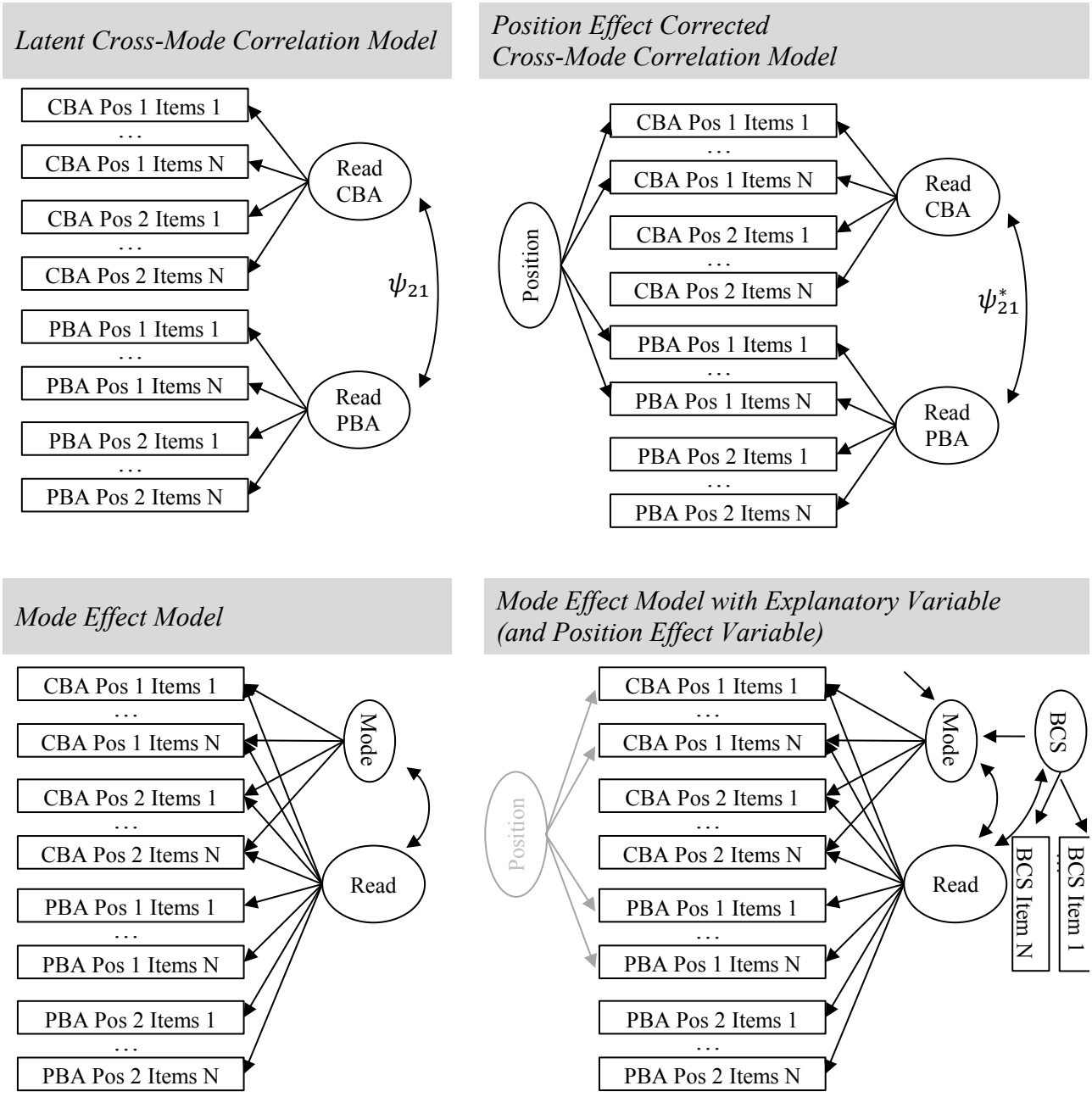


Figure 1. Schematic path diagrams of latent models for cross-mode correlation and (explained) mode effects

As it is known for ILSAs (Debeer, Buchholz, Hartig, & Janssen, 2014; Weirich, Hecht, Penk, Roppelt, & Boehme, 2016), inter-individual differences in position effects can have a strong influence (Kyllonen, 2017). As the cross mode correlation incorporates parts of the instrument administered at different positions a perfect correlation is not expected even if construct equivalence is given, until inter-individual differences in position effects are

accounted for (e.g., Bulut, Quo, & Gierl, 2017; see panel "Position Effect Corrected Cross-Mode Correlation Model" in Figure 1).

Beyond position effects, test-specific characteristics (e.g., derivations from the used latent variable model such as violations of the unidimensionality assumption of the assessed construct) might also lead to a non-perfect (latent) cross mode correlation of two parts of the instrument. Accordingly, instead of testing the deviation of the cross mode correlation from a perfect latent correlation, the evaluation of construct equivalence might alternatively compare the estimated latent correlation of two parts of the instrument across modes to the latent correlation within modes, either within PBA or within CBA. This correlation could be estimated from the mode effect study itself, if at least a subgroup of test takers answers both forms in one mode. Alternatively, the expected size of the latent correlation between the two parts of the instrument, that is the criterion, if no mode effect exists, could be estimated from available data (e.g., from a previous paper-based administration of the same instrument).

4.2 Inter-Individual Differences in the Mode Effect

Reformulating the model used to estimate the latent cross mode correlation as model with a variable for the "latent difference" between the construct measured in two different modes allows quantifying the variance that is due to the different PoTA. As Eid, Geisser and Koch (2016) summarize, different types of models were developed over the last decades to deal with the so-called *method effects* for multitrait-multimethod data, where each measurement is understood as *trait-method unit* in the tradition of Campbell and Fiske (1959). The trait (i.e., test takers reading ability) and the method (i.e., the specific PoTA used for the assessment) are confounded, as long as only one method is used. Considering modes of test administration (i.e., different PoTA) as *method* when shifting from PBA to CBA allows to differentiate and identify a latent mode effect variable that captures inter-individual differences that are due to the different PoTA.

To investigate those individual-specific mode effects empirically, we apply the model specification developed by Pohl et al. (Pohl, Steyer, & Kraus, 2008; see also Pohl & Steyer, 2010) and derive the latent mode effect variable from the difference between PBA and CBA (see panel "Mode Effect Model" in Figure 1). The definition of the individual mode effect as the difference between the two true scores relies on psychometric theory (Pohl et al., 2008) and the model can be applied to data with multiple indicators. In terms of Eid et al. (2016) assessments with different PoTA are considered as structurally different methods. Moreover, the latent trait and the mode variables are defined explicitly utilizing random assignment of test takers to modes in a balanced experimental design (see above). The mean difference between modes can be modeled and the mode effect variable is allowed to correlate with the latent trait. Finally, the model is capable to add explanatory variables to explain the inter-individual differences due to the different modes allowing to investigate fairness with respect to person-level characteristics.

4.3 Importance of Covariates

Between-subject designs such as the design used in the PISA 2015 field trail require covariates as criterion variables to judge construct equivalence, with the hypothesis that the (latent) correlation of the measured construct (e.g., reading competence) with any covariate is identical, regardless of the mode, if CBA and PBA are equivalent at the construct level. On the contrary, within-subject designs seem to be advantageous to study construct equivalence, because no additional variables are required, at least as long as no inter-individual differences in the mode effect exist (i.e., if the latent correlation between CBA and PBA is not perfect). However, if inter-individual differences between modes exist, additional person-level covariates will be necessary also for within-subject designs. Only with the help of covariates, it is possible to test hypothesis about relationships of individual differences in method effects that might indicate fairness issues and to judge the variance of the latent mode effect variable as construct-relevant or construct-irrelevant variance. As shown in Table 2, multiple

constructs can be related to inter-individual differences between modes, including i) reading competence itself (White, Kim, Chen, & Liu, 2015), ii) covariates related to computer usage and computer experience (Clariana & Wallace, 2002; White et al., 2015), and iii) covariates related to either i) or ii) such as gender (e.g., Clariana & Wallace, 2002; Leeson, 2006; Jeong, 2014). According to i), White et al. (2015) found differential effects for high-, middle- and low-performing students with respect to writing performance on computer. High-performing students performed significantly better on computer than middle and low-performing students and the authors concluded that the potential negative effect due to CBA was smaller for high-performing students. Although some authors named computer familiarity (ii) as a "*key factor associated with the test mode effect*" (Leeson, 2006, p.6), previous studies found no relation to performance differences due to the mode of assessment (e.g., Clariana & Wallace, 2002. With respect to gender (iii), results were ambiguous, and effect sizes rather small (Leeson, 2006).

5. Hypothesis

For the present study we formulate four hypotheses regarding the comparability of the reading assessment with PISA instruments between CBA and PBA. The first hypothesis refers to construct equivalence as the central criterion required for maintaining the meaning of the measured construct after introducing computer-based assessment for PISA reading competence. We try to falsify the hypothesis that due to computer-based testing additional construct-irrelevant variance is induced and thus, scores from CBA cannot be interpreted identically. In our study PISA reading items from two clusters ("Form A" and "Form B") were administered paper-based and computer-based. As the between-mode correlation can only be expected as high as the within-mode correlation in either PBA or CBA, we formulate the hypothesis with respect to the empirically estimated correlation of paper-based "Form A" and "Form B" administered in the German PISA 2009 sample as follows:

H_1: *The latent correlation of paper-based and computer-based reading comprehension is as high as the correlation of both test forms in a paper-based assessment.*

Given the selected IRT model fits as measurement model in both modes, the second hypothesis addresses the comparability of item parameters in that particular IRT model. With respect to item difficulties, a general shift due to the different PoTA is not expected. We assume that only a limited number of items show statistical significant differences in item difficulties between modes. This hypothesis is derived from the mode effects analysis in PISA 2015 (i.e., only effects of type B are expected), aiming to confirm the assumptions underlying the applied treatment.

H_2: *The item difficulties are not effected generally (as overall shift of item difficulty) and only for a subset of items the difficulties are statistically significant higher for computer-based testing compared to the paper-based test.*

The third hypothesis – also derived from the mode effect analysis in PISA 2015 – addresses the existence of inter-individual differences between modes (i.e., variance on the latent mode effect factor), as found in Model 4 (OECD, 2016, Annex A6). Derived from White et al. (2015) we expect a relationship between individual differences in the mode effect and the measured construct of reading competence. Based on hypothesis H_2 we expect the latent mode effect variable has an estimated mean of zero.

H_3: *Individual differences in the mode effect exist and are negatively correlated with reading competence, i.e., the inter-individual differences between modes are higher for students with low reading competence.*

As we expect the constructs to be equivalent (hypothesis H_1) we do not hypothesize that inter-individual differences between modes are systematically related to additional person-level covariates. From all potential covariates we focus on two covariates that seem apparently relevant: Basic computer skills as potential mediator between the true competence of test takers and their observable answering behavior and gender, as gender differences are known and consistently found in reading (OECD, 2016):

H_4: *Inter-individual differences in the mode effect can neither be explained by computer skills nor by gender.*

Finally, we expect differences between modes with respect to the observed amount of missing values. Since a missing value indicator was implemented in the computer-based test, it is expected to reduce the number of accidentally omitted responses (see Figure 2). We do not expect differences with respect to the number of not reached items, because time limits were implemented comparably between modes.

H_5: *The number of omitted items is lower for items administered on computer and the number of not reached items at the end of the test is equal for paper-based and computer-based testing.*

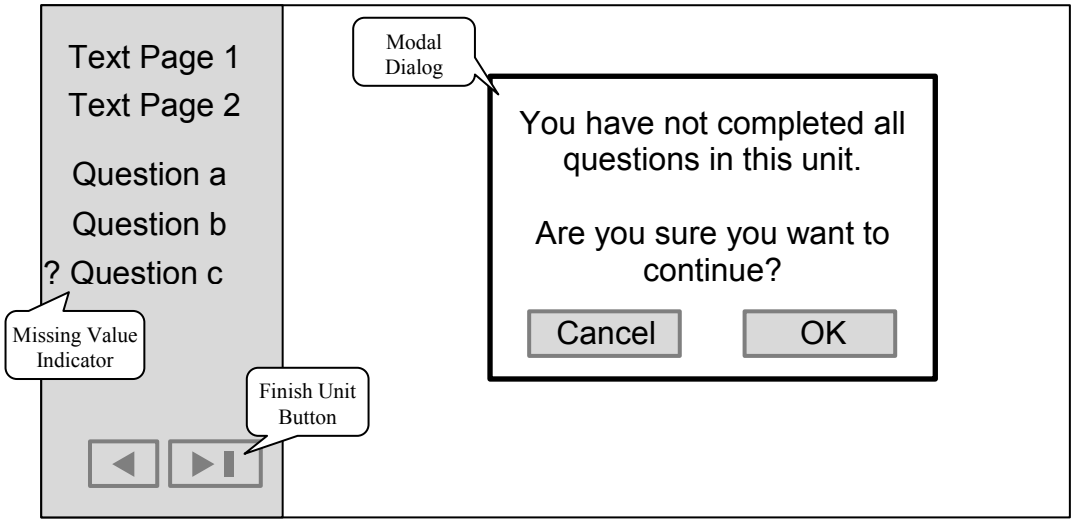


Figure 2. Illustration of the missing value indicator (question mark in front of "question c") and the modal dialog implemented as warning when exiting a unit with omitted responses

6. Method

6.1 Instruments

Reading comprehension. The items measuring reading comprehension investigated in this study were taken from the PISA 2009 reading assessment (OECD, 2009). Two intact clusters with non-overlapping items ("Form A" and "Form B") have been selected and computerized with the CBA-ItemBuilder (Roelke, 2012). The two clusters comprised five

polytomously scored items with multiple score categories and 32 dichotomously scored items, in eight units total.

Table 3 provides a verbal enumeration of the PoTA that seemed relevant for a description of the two modes. Especially the navigation (between units), the response mode for multiple choice items and short text responses as well as the availability of the missing value indicator differed between PBA and CBA.

Basic Computer Skills (BCS). Computer skills defined as generic skills to operate the graphical user interface of computers were assessed with a computer-based test comprising 20 performance-oriented tasks (Goldhammer, Naumann, & Keßel, 2013). The test was Rasch scaled and showed acceptable reliability (EAP/PV reliability 0.82). Raw items were used as indicators for a latent BCS factor predicting the latent mode effect variable (see panel “Mode Effect Model with Explanatory Variable” in Figure 1).

Category	Property	PBA Implementation	CBA Implementation
General Layout	text orientation	portrait	landscape
	font size / information per page	smaller / more	larger / less
	font family / color	identical	
Page Breaks	page breaks	according to document flow	between reading text and first question
	questions per page / screen	multiple items on page	one item on one screen
	scrolling	-	neither horizontal nor vertical
Response Process	answer selection (multiple choice items)	placing crosses with a pen	checking radio buttons or check boxes with the mouse
	answer deletion (multiple choice items)	instructed to draw erase pattern	only answer changes impossible for radio buttons
	input method (short text responses)	writing with a pen to a specified place	typing using a hardware keyboard
Navigation	within units	unrestricted page turning	links and buttons to navigate between screens
	between units	units process in self-selected sequence; possible to omit or skip pages	only screens from the current unit accessible; impossible to go back to a previous unit; impossible to skip a unit without visiting the questions
Dynamic Feedback	confirmation dialog	-	leaving a unit triggered a dialog informing about unanswered questions
	missing value indicator	-	question mark in the navigation panel that disappeared dynamically
	progress indicator	-	number of remaining questions
Hardware	description of key features	duplex printed DIN A4 paper booklets; coil binding	13” notebooks with build-in keyboard; connected mouse and de-activated touchpads

Table 3: Description of the Properties of Test Administration

6.2 Setting

Implemented as the national extension of the data collection for the PISA 2012 main assessment, all assessments took place in groups with up to 14 students in schools. A trained

test administrator supervised the test session with bring-in notebooks. The setting for the assessment was identical for students taking "Form A" or "Form B" either as PBA or CBA. Identical time limits were used for PBA and CBA and no specific feedback about the remaining time was implemented in the computerized version of the assessment. Test administration was conducted as low-stakes testing and no individual consequences were related to either good or poor performance of the test takers.

6.3 Sample

In this mode effect study 856 students (aged from 15.33 to 16.33, $M = 15.82$, $SD = 0.29$) were assessed (48.67 % female). The subset was sampled randomly from the sample of German PISA 2012 main study schools and none of the sampled schools was excluded due to technical problems. Hence, no coverage error with respect to the target population of the PISA large-scale assessment was expected. The sample contained 33.9 % students from academic track, 15.89 % were immigrants in the first or second generation, and 8.86 % reported that German is not their language at home.

6.4 Design

An experimental design with random assignment of test takers to modes was implemented. To estimate the cross mode correlation a between-subject design was supplemented by an additional within-subject component: A subset of 440 test takers answered reading items in both modes. Those students had to change the mode (i.e., switching between modes in the middle of the testing session was implemented), although some authors expect potential effects of this design element (e.g., Mazzeo & von Davier, 2013). To avoid confounding of mode and position effect the sequence of modes was balanced between subjects, resulting in the following four random equivalent groups (see Table 4): PBA (group 1), CBA (group 2), PBA-CBA (group 3) and CBA-PBA (group 4). The two parts ("Form A" and "Form B") were balanced in all groups. BCS was administered to selected rotations at second or third position. The assignment to the group was conducted at the individual level.

That is, the clustering of students in groups was not confounded with the assignment of groups.

Group	Rotation	Reading "Form A"		Reading "Form B"		BCS (CBA)
		Mode	Position	Mode	Position	Position
1a	1	PBA	1	-	-	2
	13	PBA	1	-	-	3
1b	4	-	-	PBA	1	2
	14	-	-	PBA	1	3
2a	5	CBA	1	-	-	2
	15	CBA	1	-	-	3
2b	8	-	-	CBA	1	2
	16	-	-	CBA	1	3
3a	3	PBA	1	CBA	2	3
	10	PBA	1	CBA	2	-
3b	11	CBA	2	PBA	1	3
	6	CBA	2	PBA	1	-
4a	9	CBA	1	PBA	2	3
	2	CBA	1	PBA	2	-
4b	7	PBA	2	CBA	1	3
	12	PBA	2	CBA	1	-

Note: Group 1 and 2 belong to the *between-subject* part of the design, group 3 and 4 are considered as *within-subject* design.

Table 4: Balanced Rotation Design (Two Booklets, Two Modi, Two Positions) with Random Assignment of Test Takers to Groups

6.5 Data Analysis

Latent variable models with categorical indicators were used and estimated in Mplus version 7.4 (Muthén & Muthén, 2015). In order to estimate multidimensional Partial Credit Models (PCM; Masters, 1982) a structural equation modeling (SEM) framework with continuous latent and categorical observed variables was used. A SEM approach was chosen, because i) multiple groups can be considered to capture the whole structure of the experimental design and ii), because hypotheses about differences between item parameters could be formulated using the Delta method, easily accessible in Mplus. Preceding item analyses, including item fit and graphical investigation of non-parametric item characteristic curves were conducted prior to the Mplus analysis using the packages TAM (Robitzsch,

Kiefer & Wu, 2017) and irtoys (Partchev, 2016) in R version 3.3.2 (R Core Team, 2016). Missing responses were categorized into omitted responses (i.e., items that were skipped by the test taker) and not reached items (i.e., items that were not reached due to lapsed testing time). Not reached items were ignored in the scaling, while omitted items were coded as wrong responses.

7. Results

7.1 Measurement Model

Prior to the analysis of mode effects, we analyzed the measurement model in a first step. In particular, the PCM was compared to the Generalized Partial Credit Model (GPCM; Muraki, 1992) between modes. For that purpose, different constraints on item discriminations were imposed in a model with equality constraints on item difficulties. The following four combinations of constraints on item discriminations were compared: a) discriminations constraint to be equal for all items (but different discriminations for each mode), b) discriminations unconstrained (freely estimated for each item in each mode), c) discriminations constraint to be equal between modes (but freely estimated for each item) and d) discriminations constraint between items and modes. The model with freely estimated item discriminations for each item but an equality constraint for each item that forces equal discrimination between modes (c) was selected according to the information criteria AIC and BIC (see Table 5). As a result, mode effects were investigated in the GPCM with regard to item difficulties and thresholds, while item discriminations were constraint to be equal between modes. Two items showed item misfit in PBA and CBA and thus have been excluded from the analysis. Item misfit that has been found in only one mode was ignored (five additional items with misfit only in PBA have been kept in the analysis).

Constraint on Item Discrimination	Number of Parameters	AIC	BIC
a) discriminations constrained to be equal for all items (but different discriminations for each mode)	49	23643.11	23877.32
b) discriminations unconstrained (freely estimated for each item in each mode)	117	23458.10	24017.35
c) discriminations constrained to be equal between modes (but freely estimated for each item)	83	23431.95	23823.90
d) discriminations constrained between items and modes	48	23644.25	23873.68

Note: Model c) was chosen as measurement model because of lowest AIC and BIC.

Table 5: Models with constraints on item discriminations

As a second step we inspected item misfit (evaluated on infit_t value > 2 from TAM analysis).

7.2 Construct Equivalence

Construct equivalence has been investigated by using the estimated latent correlation of PBA with CBA. Items not administered due to the rotation design were considered as missing completely at random (by design). The magnitude of the latent correlation was expected to be as high as the correlation of both test parts ("Form A" and "Form B") in a paper-based administration. As no student answered items of both forms in the same mode in the design (see Table 4), we estimated the latent correlation from PISA 2009 data (OECD, 2010) and used it as criterion for construct equivalence. A total of $N = 355$ students from the German PISA 2009 sample answered items from "Form A" and "Form B" in the PISA 2009 booklet design (Booklet 4), and $N = 2220$ students answered either items in "Form A" or "Form B". Using all data, a two-factor model was estimated in Mplus with item discriminations and thresholds estimated freely for each item (GPCM). The correlation between "Form A" and "Form B" used in our mode effect study, estimated from PISA 2009 data administered as paper-based assessment, was 0.87 ($SE = 0.03$).[2] Thus, the cross mode correlation of PBA with

CBA was expected to be not statistically different from 0.87 (H_1), if the construct was equivalent between modes.

Using the "Latent Cross-Mode Correlation Model" (Figure 1), we estimated the latent correlation ψ_{21} and tested the hypothesis that ψ_{21} is different from 0.87 using the Wald test-statistic provided by Mplus. The correlation estimated from our mode effect study between PBA and CBA was 0.86 (*SE* = 0.03), not statistically different from the expected correlation in PBA ($\chi^2 = 0.238$, $df = 1$, $p = 0.63$).[3]

7.2 Item Difficulties

Defining the difference of item difficulties between CBA and PBA as a new parameter for each item allows investigating item level mode effects that are targeted in hypothesis H_2. The item difficulty difference between modes was found to be significantly different from zero for four items (in the "Latent Cross-Mode Correlation Model", see Table 6), three items were more difficult on computer while one item was easier.

Item No.	Mode Effect (SE)
1	0.02 (0.26)
2	0.34 (0.21)
3	0.06 (0.23)
4	0.80 (0.27) **
5	0.69 (0.26) **
6	0.03 (0.22)
7	0.17 (0.22)
8	0.50 (0.25)
9	0.39 (0.19)
10	-0.04 (0.19)
11	-0.27 (0.21)
12	-0.18 (0.21)
13	-0.35 (0.34)
14	-0.11 (0.25)
15	-0.75 (0.24) **
16	0.03 (0.18)
17	-0.43 (0.36)
18	-0.22 (0.24)
19	0.10 (0.23)
20	0.05 (0.29)
21	-0.1 (0.18)
22	0.25 (0.26)
23	0.01 (0.19)
24	0.00 (0.19)
25	-0.05 (0.28)
26	-0.26 (0.21)
27	-0.03 (0.32)
28	-0.03 (0.17)
29	-0.24 (0.21)
30	0.44 (0.21) *
31	0.00 (0.28)
32	0.12 (0.20)
33	-0.18 (0.20)
34	0.18 (0.18)
35	0.40 (0.20)

Note: * p <.05, ** p< .001. A positive effect means that this item was more difficult on computer while a negative value indicated that this item was more difficult in PBA.

Table 6: Mode Effects on the Item Level

Constraining all differences in item difficulties to one common parameter (homogenous shift) resulted in an estimated mode effect that is not different from zero (0.01, $SE = 0.05$, $p = 0.79$). That means there was no homogeneous shift for all item difficulties that is different from zero when assessed on computer. However, according to the information criteria the

constrained model with 83 free parameters is slightly preferred in terms of lower AIC (23433.87 vs. 23436.61) and lower BIC (23436.60 vs. 23995.86) over the unconstrained model with 117 estimated parameters.

7.3 Inter-Individual Differences

For investigating inter-individual differences due to the mode change an additional latent mode effect factor was included in the Mplus model. In line with our hypothesis H_2 the mean of the latent mode effect variable was not statistically different from zero (-0.04, *SE* = 0.05, $p = 0.44$). This means that on average the computer-based test was of the same difficulty as the paper-based test. However, inter-individual differences in the mode effect were estimated as variance on this mode factor (0.36, *SE* = 0.09, $p < 0.01$), meaning that for some test takers the computer-based reading assessment was more difficult than for others. Moreover, the mode factor was negatively correlated (-0.41, *SE* = 0.08, $p < 0.01$) with the latent ability of reading comprehension (H_3). This correlation indicates that the mode effect is positive for students with lower reading ability (i.e., the computer-based test tends to be more difficult compared to the paper-based test) and negative for students with high reading ability (i.e., the computer-based test tends to be easier compared to the paper-based test). Note that the variance on the mode effect vanishes, when the position effect is included (mode effect mean: -0.04, *SE* = 0.08, $p = 0.44$; mode effect variance: 0.19, *SE* = 0.12, $p = 0.13$; correlation between mode effect and reading: -0.37, *SE* = 0.12, $p < 0.01$), although the variance of the position effect variable is not statistically different from zero (0.144, *SE* = 0.13, $p = 0.25$) while the mean was (0.65, *SE* = 0.05, $p = 0.03$).

7.4 Covariates

To further investigate whether the observed inter-individual differences between modes pose a threat to the construct equivalence we tried to explain the variance of the latent mode effect variable with covariates. For that purpose, we estimated the "*Mode Effect Model with Explanatory Variable"* (see Figure 1) with the covariates separately (H_4). To test the

relationship of the mode effect with basic computer skills, we added BCS as predictor of the latent mode effect variable. The analysis yielded no significant effect for BCS (-0.02, *SE* = 0.10, p = 0.88). That is, the mode effect was not linearly related to computer skills as measured with the performance based BCS test. Note that BCS was correlated with reading (0.72, *SE* = 0.03, $p < 0.01$). This correlation of BCS and reading increased (0.79, *SE* = 0.04, $p < 0.01$), when the position effect was included. The linear regression of the mode effect variable on the manifest indicator for gender revealed a non-significant effect (0.10, *SE* = 0.08, p = 0.22) as well as a reduction of the latent correlation between the mode effect variable and reading (-0.24, *SE* = 0.06, $p < 0.01$). The correlation was further reduced to a non-significant value (-0.14, *SE* = 0.09, p = 0.11) when the position effect was included in the model, while the regression coefficient for mode on gender was still not different from zero (0.07, *SE* = 0.08, p = 0.34).

7.5 Missing Values

Finally, with respect to the number of missing values, we found a statistical difference between both modes with respect to omitted responses and no statistical difference with respect to not reached items confirming our hypothesis H_5. Overall 5.15 percent of the items were omitted in PBA and 3.48 percent in CBA. In PBA 2.52 percent of the items were not reached at the end of the test while in CBA 2.05 percent were not reached.

8. Discussion

8.1 Main findings

In this paper the equivalence of paper-based and computer-based assessment of reading comprehension was investigated. From the actually applied treatment of mode effects in PISA 2015 we derived the need to investigate construct equivalence of computer-based measured reading comprehension as central equivalence criterion. As prerequisite for the following latent variable modeling we estimated a GPCM as measurement model. This was in line with the scaling of PISA 2015 data where the GPCM was used, although we had to exclude one

misfitting item. With respect to construct equivalence the main result of this study is that the cross-mode correlation of PBA and CBA, estimated as latent correlation between two subsets of items ("Form A" and "Form B") was not statistically different from the correlation of both test forms estimated from the paper-based administration in PISA 2009.[4]

Inter-individual differences on an included latent mode effect factor could neither be explained by basic computer skills nor gender. This is an important finding with respect to fairness of computer-based testing. Due to the transition from PBA to CBA, students with lower computer skills were not disadvantaged nor led the transition to a discrimination with respect to gender. However, the latent mode effect variable was negatively correlated with reading. This indicated that the mode effect was higher for students with lower reading ability, but zero on average. The results give no reason to assume that inter-individual differences in the mode effect are the result of gender differences or differences in computer skills. However, further research is needed to disentangle the interplay between the inter-individual differences in the position effect, gender and the negative correlation of the latent mode effect variable and reading.

No empirical evidence was found that the mode effect on item difficulties is on average different from zero for the investigated subset of items. We found a small number of item-level mode effects that were statistically different from zero. However, as we did multiple significance tests uncorrected for the multiple testing we refrain from overemphasizing these item-level mode effects.

With respect to the number of missing values we found a small tendency to work faster in computer-mode that is in line with Bodmann and Robinson (2004). However, the difference in the number of not reached items at the end of the test was not statistically significant. According to the number of omitted responses we found significant less omitted responses in CBA. As a missing value indicator was implemented on computer this result may lead to the conclusion, that the indicator diminishes the number of items skipped accidentally.

Further research using log-data from computer-based testing is necessary to investigate, if this reduced number of missing values is truly an effect of the presented dialog.

8.2 Limitations

Even though our findings regarding the construct equivalence might help to dispel doubts about mode effects in the PISA trend estimates, our study is limited in various ways. Firstly, PISA is an international large-scale assessment, conducted in many countries. Hence, for the interpretation of mode effects in PISA, estimates at the country level (i.e., mode-by-country interactions) must be considered. However, we cannot provide evidence that the results presented in this paper can be generalized for different countries as our add-on study was administered to PISA 2012 in Germany only. Secondly, as the CBA implementation investigated in this study differed with respect to some PoTA from the CBA implementation used in PISA 2015, the generalizability of our findings must be verified in further research. This in particular true with respect to our findings regarding the differences in the amount of missing values due to the missing value indicator included in our CBA implementation. Thirdly, a general limitation of this study is the small sample size. This might have led to a reduced power of the statistical analyses and prevented us from more detailed analyses of mode effects in performance subgroups. Further research on mode effects should necessarily use sufficiently large sample sizes in within-subject designs to ensure power of analyses especially for estimating multidimensional models (including position effects) as well as latent mode effect variables and should elaborate on the practical consequences of the relationship of the inter-individual differences between modes and the comparability of competences measured in PBA and CBA. Upcoming mode effect studies should investigate the relationship of inter-individual differences in the mode effect to further covariates that might be theoretically related to structural changes (see Table 2). For instance, time-related differences between modes should be taken into account, as it is known that reading

efficiency and reading speed are related to reading comprehension and differ between modes (Kerr & Symons, 2006).

8.3 Conclusion

This study showed that the transition to computer-based testing did not lead to "big changes". When divided into two parts, the correlation of the parts between computer-based and paper-based assessed reading competence is comparable to the correlation of the two parts in one mode. We found no evidence that the construct should not be considered as equivalent between modes. As another major finding with respect to fairness, neither gender nor a performance based measure of computer skills predicted inter-individual differences between computer-based and paper-based assessment. Only some items showed a statistical significant difference in item difficulty while most of the items were found to be invariant (individually and on average) in terms of difficulty and thus could be used as link items to create a common metric for comparing PBA scores with CBA scores.

9. References

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

Bodmann, S. M., & Robinson, D. H. (2004). Speed and Performance Differences among Computer-Based and Paper-Pencil Tests. *Journal of Educational Computing Research*, 31(1), 51–60.

Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The Transition to Computer-Based Testing in Large-Scale Assessments: Investigating (Partial) Measurement Invariance between Modes. *Psychological Test and Assessment Modeling*, 58(4), 587–606.

Bulut, O., Quo, Q., & Gierl, M. J. (2017). A structural equation modeling approach for examining position effects in large-scale assessments. *Large-Scale Assessments in Education*, 5(1), 8.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance. *Applied Measurement in Education*, 16(3), 191–205.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.

Clariana, R., & Wallace, P. (2002). Paper–based versus computer–based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*, 33(5), 593–602.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, 39(6), 502–523.

Eid, M., Geiser, C., & Koch, T. (2016). Measuring method effects: From traditional to design-oriented approaches. *Current Directions in Psychological Science*, 25(4), 275–280.

Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing Individual Differences in Basic Computer Skills: Psychometric Characteristics of an Interactive Performance Measure. *European Journal of Psychological Assessment*, 29(4), 263–275.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical Guidelines for Assessing Computerized Adaptive Tests. *Journal of Educational Measurement*, 21(4), 347–360.

Jerrim, J. (2016). PISA 2012: how do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice*, 23(4), 495–518.

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology*, 33(4), 410–422.

Kerr, M.A., & Symons, S.E. (2006). Computerized Presentation of Text: Effects on Children's Reading of Informational Material. *Reading and Writing*, 19(1), 1–19.

Kim, D.-H., & Huynh, H. (2008). Computer-Based and Paper-and-Pencil Administration Mode Effects on a Statewide End-of-Course English Test. *Educational and Psychological Measurement*, 68(4), 554–570.

Klieme, E. (2016). TIMSS 2015 and PISA 2015: How are they related on the country level? (DIPF Working Paper) Retrieved from DIPF website https://www.dipf.de/de/forschung/publikationen/pdf-publikationen/Klieme_TIMSS2015andPISA2015.pdf

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking*. New York, NY: Springer New York.

Komatsu, H., & Rappleye, J. (2017). Did the shift to computer-based testing in PISA 2015 affect reading scores? A View from East Asia. *Compare: A Journal of Comparative and International Education*, 47(4), 616–623.

Kroehne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft*, 14(S2), 169–186.

Kyllonen, P. C. (2017). Socio-emotional and Self-management Variables in Learning and Assessment. In: Rupp, A. A., Leighton, J. P. (Eds.), *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (pp. 174–197). Wiley & Son.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, 6(1), 1–24.

Lottridge, S. M., Nicewander, W. A., Mitzel, H. C., & Metrics, P. (2010). Summary of the Online Comparability Studies for One State's End-of-course Program (Section 2: Studies of Comparability Methods). In: P. C. Winter (Ed.), *Evaluating the Comparability of Scores from Achievement Test Variations* (pp.13–32), Council of Chief State School Officers, Washington, DC.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, 58, 61–68.

Mazzeo, J. & Davier, M. von (2008). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. Education Working Papers EDU/PISA/GB (2008), 28, 23–24.

Mazzeo, J., & von Davier, M. (2013). Linking scales in international large-scale assessments. In: Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.) *Handbook of international*

large-scale assessment: Background, technical issues, and methods of data analysis, (pp. 229–258), CRC Press: New York

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education*, 39(3), 299–312.

Mead, A. D., & Drasgow, F. (1993). Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. *Psychological Bulletin*, 114(3), 449–458.

Meredith, W. (1993), Measurement invariance, factor analysis and factorial invariance, *Psychometrika*, 58(4), 525–43.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *ETS Research Report Series,* (1), 1–30.

Muthén, L.K. & Muthén, B.O. (1998-2015). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics*, 51(9), 1352–1375.

OECD. (2016). PISA 2015 Results (Volume I). OECD Publishing.

OECD. (2015). Adults, Computers and Problem Solving: What's the Problem? OECD, Publishing.

OECD. (2010). *PISA 2009 Results: Executive Summary*. OECD Publishing.

OECD. (2009). PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science. OECD Publishing.

Partchev, I. (2016). irtoys: A Collection of Functions Related to Item Response Theory (IRT). R package version 0.2.0. https://CRAN.R-project.org/package=irtoys

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large

scale state assessment program. *The Journal of Technology, Learning and Assessment*, 3(6).

Pohl, S., & Steyer, R. (2010). Modeling Common Traits and Method Effects in Multitrait-Multimethod Analysis. *Multivariate Behavioral Research*, 45(1), 45–72.

Pohl, S., Steyer, R., & Kraus, K. (2008). Modelling method effects as individual causal effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(1), 41–63.

Pommerich, M. (2004). Developing Computerized Versions of Paper-and-Pencil Tests: Mode Effects for Passage-Based Tests. *The Journal of Technology, Learning, and Assessment*, 2 (6), 3–44.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Roelke, H. (2012). The ItemBuilder: A Graphical Authoring System for Complex Item Development (Vol. 2012, pp. 344–353). Presented at the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education.

Robitzsch, A., Kiefer, T., & Wu, M. (2017). TAM: Test analysis modules. R package version 2.1-43. https://CRAN.R-project.org/package=TAM

Robitzsch, A., Luedtke, O., Koeller, O., Kroehne, U., Goldhammer, F., & Heine, J. H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. *Diagnostica*, 63(2), 148–165.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of Reading and Listening Comprehension Across Test Media. *Educational and Psychological Measurement*, 71(5), 849-869.

Vinsonhaler, J. F., Molineaux. J. E., & Rodgers, B. G. (l968). An experimental study of computer-aided testing. In H. H. Harman, C. E. Helm, & D. E. Loye (Eds.), *Computer-Assisted Testing Conference Proceedings*, November 1966, Princeton, NJ: Educational Testing Service.

von Davier, M. (2005). mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]. Princeton, NJ: ETS.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K 12 Reading Assessments: A Meta-Analysis of Testing Mode Effects. *Educational and Psychological Measurement*, 68(1), 5–24.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Boehme, K. (2016). Item Position Effects Are Moderated by Changes in Test-Taking Effort. *Applied Psychological Measurement,* 41(2), 115–129.

White, S., Kim, Y., Chen, J., & Liu, F. (2015). Performance of fourth-grade students in the 2012 NAEP computer-based writing pilot assessment: Scores, text length, and use of editing tools (NCES 2015–119). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Yamamoto, K. (2012). *Outgrowing the Mode Effect Study of Paper and Computer Based Testing*. Retrieved from http://www.umdcipe.org/conferences/EducationEvaluationItaly/COMPLETE_PAPERS/Yamamoto/YAMAMOTO.pdf

Footnotes

[1] Even though Model 4 in the summarized treatment of mode effects in PISA 2015 (OECD, 2016, Annex A6) includes a second latent variable that is assumed to be uncorrelated to the latent ability, the comparison between Model 3 and 4 with respect to information criteria and model parsimony is at the very most a weak attempt to falsify the hypothesis of construct equivalence.

[2] The correlation estimated from PISA 2009 data was not corrected for position effects because only Booklet 4 contained all items of "Form A" and "Form B". Hence, it was not possible to identify a model that incorporates inter-individual differences in the position effect.

[3] We also estimated the latent cross mode correlation corrected for position effects ψ_{21}^{*} (see "Position Effect Corrected Cross-Mode Correlation Model" in Figure 1) and tested the hypothesis that this correlation differs significantly from a perfect latent correlation. The cross mode correlation corrected for position effects was estimated as 0.92 (SE = 0.03), not statistically different from 1.0 ($\chi^2 = 3.12$, df = 1, p = 0.08). However, in this model neither the variance nor the mean of the latent position effect model reached statistical significance.

[4] We also found that, taking the position effect into account, the cross-mode correlation was not statistical significant different from one. However, this might be the result of a lack of statistical power as the model including mode and position effect becomes three-dimensional. Therefore, this result might be preliminary and needs to be interpreted cautiously since the variance of the position effect did not reach statistical significance while a latent mode effect variable showed variance significantly different from zero.

Anhang C

Beitrag III: Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (*submitted, Educational and Psychological Measurement*). What makes the difference? The Impact of Item Properties on Mode Effects in Reading Assessments.

What makes the difference? The Impact of Item Properties on Mode Effects in Reading Assessments

Sarah Buerger[1], Ulf Kroehne[1], Carmen Koehler[1], Frank Goldhammer[1,2]

[1]German Institute for International Educational Research (DIPF)

[2]Centre for International Student Assessment (ZIB)

Author Note

Sarah Buerger, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany; Ulf Kroehne, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany; Carmen Koehler, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany; Frank Goldhammer, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany.

Correspondence concerning this article should be addressed to Sarah Buerger, German Institute for International Educational Research (DIPF), Solmsstraße 73-75, 60486 Frankfurt am Main, Germany.

E-mail: buerger@dipf.de

Abstract

As part of the transition from paper-based assessment (PBA) to computer-based assessment (CBA), mode effect studies are required to investigate the comparability of test scores across modes. In the National Educational Panel Study (NEPS), experimental bridge studies were conducted to investigate psychometric differences between modes. In the present study, the cross-mode equivalence of a NEPS reading test was examined. In particular, the investigation sought to determine whether item-level mode effects can be explained by item properties such as the response format, the need to navigate through a multi-page text, and the need to navigate between the text and related items. Additionally, a missing value indicator was implemented on the computer to avoid accidental skipping of items. The results showed that splitting texts between multiple screens, and thus requiring navigation between pages, did not affect comparability. However, item difficulty was increased in CBA when items in the first and second position of a unit were not presented on the same double-page as in PBA. Regarding response formats, assignment tasks on the computer requiring the use of combo boxes were more difficult than on paper, while no difference was found for (complex) multiple choice items. As expected, the omission rate was smaller in CBA than in PBA, highlighting the fact that the transition to computers provides an opportunity to deliberately implement new desirable properties.

Keywords: mode effect, equivalence, administration mode, measurement invariance, item properties

Introduction

Due to improvements in and the increasing availability of technology in recent decades, computer based assessments (CBA) have become increasingly common in large-scale assessments (e.g., Programme for International Student Assessment [PISA]; OECD, 2016) and longitudinal large-scale assessments (e.g., National Educational Panel Study [NEPS]; Blossfeld, Roßbach, & von Maurice, 2011). The ongoing transition from paper-based assessments (PBA) to CBA raises the question of equivalence and valid comparisons between test results obtained from paper-based and computer-based instruments. Mode effects, defined as non-equivalence in psychometric item and/or test properties, might affect test results and produce scores that are not directly comparable. This means that two persons with the same underlying proficiency in the measured construct might obtain different test scores in the CBA and the PBA versions (Van den Noortgate & De Boeck, 2005). Without empirical evidence about their equivalence, scores from computer- and paper-based instruments cannot be used interchangeably, because the existence of mode effects cannot be ruled out. Psychometric properties of the test that might be affected by a mode change are the represented construct (in terms of introduced construct-irrelevant variance due to computer-based testing) as well as the reliability of the test and item parameters (e.g., difficulty and discrimination). In a mode effect analysis, those properties can be used as criteria for evaluating equivalence between modes. If the PBA and the CBA versions are completed by random equivalent groups and if there is equivalence in the underlying construct, scores can still be used interchangeably. Specifically, by imposing equality constraints on the mean and variance of the person parameter distribution between groups, item parameters can be estimated per mode and thus scores can be adjusted for mode effects. However, if construct-irrelevant components are measured with CBA, such as ICT literacy, and individual differences in that component have an effect on test scores, scores cannot be used interchangeably even if mode-specific item parameters are estimated. The appropriateness of

comparing test scores across modes has to be investigated in equivalence studies, and the findings of these studies have to be documented as required by testing standards (e.g., AERA, APA, & NCME, 2014; American Psychological Association, Committee on Professional Standards [COPS] and Committee on Psychological Tests and Assessments [CPTA], 1986; Association of Test Publishers [ATP], 2000; International Test Commission [ITC], 2005).

Large-scale assessments such as PISA (OECD, 2016; OECD, 2014) and the Programme for the Assessment of Adult Competencies (PIAAC; OECD, 2015; OECD, 2013) recently changed their administration mode from PBA to CBA. Thus, evidence of comparability between those different administration modes is needed to ensure stable trend measures (Mazzeo & von Davier, 2008; OECD, 2016, Annex 6). The challenge of a mode switch is not only important for large-scale assessments such as PISA and PIAAC but also for longitudinal large-scale studies such as NEPS (Blossfeld et al., 2011). The present study investigates mode effects in a reading test from an experimental bridge study in NEPS that was conducted to investigate psychometric differences between modes. NEPS is an extensive longitudinal study that applies a large-scale multi-cohort sequence design with competence testing in the domains of mathematics, reading, science, and ICT literacy (Artelt, Weinert, & Carstensen, 2013). Thus, the comparability of repeated assessments is an important prerequisite for both comparisons between cohorts and deriving change scores within cohorts. Hence, the change to CBA in NEPS requires mode effect studies to investigate the comparability of competence measurements across modes.

In the present study, the cross-mode equivalence of a NEPS reading competence test (Gehrer, Zimmermann, Artelt, & Weinert, 2013) is investigated. Our focus is on item properties as potential sources of mode effects, as mode effects are especially likely to arise in reading tests due to their long reading passages and related navigation possibilities in CBAs (Pommerich, 2004; Wang, Jiao, Young, Brooks, & Olson, 2008). Thus, we aim to not only describe a general or item-specific mode effect, but also to tie it to properties of the

computerized test version. This provides the basis for explaining and predicting mode effects. In the next section, properties of test administration that might differ between PBA and CBA are introduced, followed by a presentation of equivalence criteria used in empirical attempts to falsify the hypothesis of equivalence.

Properties of Test Administration

Mode effects can appear on the overall test level as a homogeneous effect, which is similar for all items, or on the item level if the mode effect interacts with items or their properties (Green, Bock, Humphreys, Linn, & Reckase, 1984). With the transition from PBA to CBA among existing instruments, various properties of test administration (i.e. the way items are presented, answers are recorded, orientation within the test, and navigation between items) have been changed, which may affect the test-taking process (Kroehne & Martens, 2011; Pommerich, 2004), even if the computerized version was created to match the paper-based version as closely as possible. Hence, differences between modes might be caused not only by a single measurement property but also by an aggregation of properties that affect the comparability of tests.

A meta-analysis by Wang et al. (2008) analyzing studies conducted between 1980 and 2005 showed that results from previous mode effect studies on reading assessment are quite heterogeneous. The different methods used to compare CBA and PBA as well as differences in sample size, study design and computer software led to inconsistent results regarding cross-mode equivalence (Wang et al., 2008). More recent studies conducted after 2005 still showed heterogeneous results, which also varied over domains (e.g. Bennett et al., 2008; Coniam, 2006; Kim & Huynh, 2008; Pommerich, 2007; Robitzsch et al., 2016; Schroeders & Wilhelm, 2011). The conclusion is that mode effects cannot be predicted or ruled out in general, especially in reading assessments, but need to be explicitly tested whenever a previously paper-based test is implemented on the computer and the results from both test versions are to be comparable (Pommerich, 2004; Schroeders & Wilhelm, 2011; Wang et al., 2008). In

addition, detailed analyses on the item level examining the effect of test administration properties are of utmost interest in terms of identifying potential sources of mode effects (Pommerich, 2004). Drawing on empirical evidence from previous studies, the following sections briefly review those properties which are particularly relevant for the present study.

Test and Item Layout

Item layout may differ between tests on the computer and on paper because of the reduced screen size and different page orientation, which has an effect on the size and placement of text (e.g., column and line breaks), images and graphics (Pommerich & Burden, 2000; Schroeders & Wilhelm, 2010)., The text stimulus is often presented on multiple pages in CBA reading tests to avoid scrolling (in contrast to a single or double page for PBAs). Such page breaks require switching back and forth between pages, which might affect the psychometrical characteristics of the item.

Navigation

Reading tests are often structured by units, that is, a reading text (stimulus) is presented together with multiple items (questions) to be answered using information provided in the text. In CBA, navigation often is required because not all of the information for an item can be presented on one screen. This makes reading tests more prone to mode effects than mathematics tests, for example (Pommerich, 2004).

Regarding item review, backward navigation between and within units is typically not restricted in PBA, and test takers are allowed to temporarily skip items or return to previous items in order to change their answer or fill in missing responses at any time. In contrast to PBA, unrestricted navigation and item review in CBA is usually only allowed within units (Vispoel, 2000). Although there is a technical difference between modes, previous studies have shown that test takers' navigation behavior within PBA booklets is not substantially different from the restricted navigation in CBA (Kroehne, Roelke, Kuger, Goldhammer, &

Klieme, 2016) and that preventing item review does not affect average scores between modes (Vispoel, 2000).

Item Response Format

The response format of items may differ between PBA and CBA, and may be used differently by test takers (Sireci & Zenisky, 2006). In this regard, previous studies have shown that the complexity of the response format has to be considered, which can be defined as the required "number and type of examinee interactions within a given task or item" (Parshall, Spray, Kalohn, & Davey, 2002, p.9). For instance, it has been shown that computerized multiple choice items are less prone to mode effects because they are of lower item complexity (Bennett et al., 2008; Bodmann & Robinson, 2004; Parshall et al., 2002). In contrast, assignment tasks are of higher complexity. They are typically used in reading tests, for instance, when test takers are required to assign a set of given headings to paragraphs in the text. Assignment tasks can be computerized using what are known as combo boxes (or drop-down boxes; see Figure 1). A drop-down menu has to be opened and an entry has to be selected, whereas on paper responses are given by writing down a character in the response field. Combo box items in CBA have been found to be more difficult than assignment tasks on paper tests (Heerwegh & Loosveldt, 2002).

Insert Figure 1 about here

CBA Missing Value Indicator

As computers can easily change the visual presentation of items depending on response behavior, studies comparing administration modes lend themselves to investigating the (desirable) effect of these new possibilities. For example, CBA allows for the implementation of a missing value indicator, an automatic feedback mechanism that informs test takers about omitted items at the end of a unit (see Figure 2). They then can decide whether to go back and provide a response for those items or skip them and go on to the next unit. This reduces the number of items omitted accidentally (Wang et al., 2005). In addition,

within a unit, visual feedback about whether an item has been completed is possible, e.g. by providing an overview of all items in a unit and highlighting those that have not yet been answered (see Figure 2). Note that in the case of such feedback, the perspective on the sources of mode effects is somewhat changed in that the focus is not on how item and test properties are changed by computerization (with respect to, e.g., layout, navigation, response format), but on adding a desirable feature that was not available before.

Insert Figure 2 about here

The Present Study

The current study investigates the equivalence of the PBA and CBA versions of a NEPS reading competence test (Gehrer et al., 2013). It extends previous research by presenting a practical application of a multiple-group IRT model in which mode effects are represented as group differences (for conceptual details see Buerger, Kroehne, & Goldhammer, 2016). This model allows in-depth analyses of mode effects on the test and item levels using psychometric criteria to be conducted to ensure the required level of equivalence between tests. An advantageous model extension used in the present study refers to the inclusion of item properties, which then allows for an analysis of their effect on mode differences. Thus, this study aims to thoroughly test mode effects by defining explicit criteria for falsifying the hypothesis of equivalence and by investigating whether item properties can explain mode effects on the item level. The item properties considered were the item response format, the number of text pages and whether navigation is required between text stimulus and items (questions). Furthermore, missing value rates between modes are compared to investigate the effect of a missing value indicator implemented in CBA. In the following sections, the equivalence criteria and related hypotheses are presented.

Measurement Model and Construct Equivalence

Our first step was to determine a measurement model that fits the data from both modes. This model formed the basis of the mode effect analysis because it can provide

parameters in which mode differences can be described adequately (e.g., differences in item difficulties and item discriminations). Determining the measurement model for a construct also raises the question of whether the same latent construct is assessed in both modes (AERA, APA, & NCME, 2014; Huff & Sireci, 2001; ITC, 2005; Parshall et al., 2002; Penfield & Camilli, 2007; Puhan, Boughton, & Kim, 2007; Russell, Goldberg, & O'Connor, 2003). Construct equivalence across modes can be investigated by comparing the correlations of mode-specific latent variables with external variables (Buerger et al., 2016). Thus, to provide evidence for construct equivalence, we investigated whether the relationship between reading competence and other variables (a paper-based reading test for lower grades and a test of basic computer skills) is comparable across modes (AERA, APA, & NCME, 2014). Although this question has rarely been investigated, the research that is available has found that the construct did not change when switching to the computer (Kim & Huynh, 2008; Neuman & Baydoun, 1998). Accordingly, the first hypothesis was as follows:

H_1: The latent correlations of the reading test with external variables obtained from a) a paper-based reading test for lower grades and b) a test of basic computer skills are equal across modes.

Reliability

The comparability of test reliabilities is important when the mode of a longitudinal study is changed, and equating or linking across modes is required. If tests are not equally reliable, the results of linking methods are affected (Dorans, Moses, & Eignor, 2010). Thus, the criterion of equal reliabilities needs to be tested (AERA, APA, & NCME, 2014; Holland & Dorans, 2006; ITC, 2005; Kolen & Brennan, 2004). Reliabilities may differ between modes if, for example, the discrimination for all items on the computer is shifted in one direction compared to PBA. Previous studies showed no or at least no substantial differences (Kim & Huynh, 2008; Poggio, Glasnapp, Yang, & Poggio, 2005; Wang, 2004). In the next hypothesis, reliabilities were expected not to be significantly different between modes:

H_2: The reliabilities of the PBA and CBA versions of the reading competence test are equal.

Item Parameters

Mode effects on the item level were investigated by comparing item discriminations as well as item difficulties (for dichotomous items) and thresholds (for polytomous items). Therefore, three hypotheses on item parameters were formulated:

$H_{3.1}$: Item discriminations are equal between the computer-based test and the paper-based test.

Although we expected item discriminations to be equal, item difficulties might differ, since some properties of the computer-based reading test were expected to increase item difficulty (e.g., more complex response format).

$H_{3.2}$: The sum of all item differences in difficulty between modes is different from zero, since mode effects are expected to be in the direction of higher difficulty on the computer. Hence, an overall mode effect will be apparent on the test level and differences between items will not cancel each other out.

Given the results of previous studies that found mode effects varying across items with different properties (e.g., items with a more complex response format on the computer are more difficult; Bennett et al., 2008; Bodmann & Robinson, 2004; Parshall et al., 2002), we did not assume a strictly homogeneous mode effect, but that the mode effect would vary across items.

$H_{3.3}$: The difference in difficulties (and thresholds) across the two modes (PBA vs. CBA) differs between items; that is, there is no homogenous mode effect but rather a mode effect on the item level.

Item Properties

The item properties that we assumed would explain the item level mode effects on difficulty were the item response format, the number of text pages, and whether navigation between the reading text and item (question) was required.

Three different response formats were used in the NEPS reading test: multiple choice format, complex multiple choice format (multiple statements that required decisions between wrong / right and disagree / agree), and a combo box format for assignment tasks (see Pohl & Carstensen, 2012). Combo box items (see Figure 1) on the computer have a higher degree of complexity than assignment task items on paper because these items require a) opening a drop down menu using the pointing device (mouse), b) scrolling down a presented list if not all entries are immediately visible and c) selecting an entry[1]. Table 1 displays the number of items applying each response format (Set 1-3). Three hypotheses addressing the response format were derived:

$H_{4.1}$: Assignment tasks are more difficult on the computer than on paper because using combo boxes on the computer entails a greater risk of error than writing a character.

$H_{4.2}$: Based on previous findings, no mode effects are expected for multiple choice items.

$H_{4.3}$: Items with a complex multiple choice response format should not differ in difficulty between modes, since they can be handled similarly.

Insert Table 1 about here

In the NEPS computer test, the reading texts for each unit ranged from one to two screens, whereas in the paper condition the text for each unit was presented on only one page. On the computer, navigation between screens was required for texts split between multiple screens (see Table 1, Set 4). The first two items of a unit on paper were presented on the same page as the reading text, whereas on the computer this was not possible, meaning that test takers always had to navigate from the text to the first item and then the second, and, if

necessary, back to the text again (see Table 1, Set 5). Consequently, although turning pages and navigation between screens was required in both modes, no navigation was necessary to read the full text and answer the first two items of each unit in the paper-based test. We expected that the additional navigation requirements in CBA increase item difficulty, because less information is visible at one time compared to PBA (Pommerich, 2004). Thus, two further hypotheses were formulated:

$H_{4.4}$: Items with text stimuli that are presented without page breaks on paper but on more than one computer screen are more difficult on the computer.

$H_{4.5}$: Items in the first and second position of a unit that are presented on the same double page on paper but on separate screens on the computer are more difficult on the computer, because navigation from the text to those items is required.

Navigation in the computerized NEPS reading test was restricted to within-unit navigation, whereas navigation was not restricted at all on paper. Although test-taking time cannot be consumed by navigation within the test, it is not expected that this led to substantial differences in items not reached at the end of the CBA compared to the PBA, because test-takers' navigation behavior in PBA is supposed to be similar to restricted navigation in CBA (Kroehne et al., 2016).

$H_{5.1}$: The number of not-reached items at the end of the test is equal for the computer-based test and the paper-based test.

A reminder about missing answers when finishing a unit was given in the computer-based test so that the test takers could decide to either go back and respond to missing items or finish the unit without answering them. Additionally, the navigation bar listed all items within a unit. A question mark displayed in front of each item name disappeared when a response was given for that item (see Figure 2). Because of these indicators of non-completed

items, we expected a lower number of omitted items in the computer-based test compared to the paper-based test.

$H_{5.2}$: The number of omitted items differs between modes, with fewer omitted items for the computer-based test.

Method

NEPS is a German longitudinal study following a multicohort sequence design that aims to investigate educational processes over the life course. Instruments for competence testing in the NEPS are developed for specific grades and age groups. This mode effect study was conducted specifically to prepare for the mode change in the NEPS main studies and is thus not part of the panel.

Subjects. The sample consisted of $N = 1163$ students in the seventh grade. Classes of up to $N = 24$ students were tested in schools from different German school tracks. The sample included 593 (51 %) female and 570 (49 %) male students.

Research design. The students within classes were randomly assigned to mode groups in a between-subject design. Thus, half of the students took the paper-based test and half of the students took the computer-based test. The students also completed further tests and a questionnaire.

Instruments. We investigated mode effects in a reading test (Gehrer et al., 2013) developed for the NEPS main study targeting students in Grade 7 (administered to stating cohort 3 in 2012; see Artelt et al., 2013). The reading competence test for Grade 7 had two forms that varied in difficulty (READ 7A and READ 7B). Each form consisted of five units, three of which were included in both forms. The results of READ 7A and READ 7B are the outcome measure of reading competence for the sample of 7th graders. Because of the common units in READ 7A and READ 7B, the data from both samples were combined into one data set (READ 7) including all units from READ 7A and READ 7B. In sum, 42 items were analyzed for mode effects. Due to the 4 units with no overlap, the data set contained

missing values that were missing by design. All students also took an additional paper-based reading test for Grade 5 (READ 5; see Table 2). The computerization of the NEPS reading test was conducted using the CBA ItemBuilder (Roelke, 2012).

Furthermore, the students filled out a questionnaire about their experience with and use of computers as well as some questions about their reading behavior. In addition, they took a test that measures basic computer skills that involved simulated computer environments (BCS, further development of Goldhammer, Naumann, & Keßel, 2013).

Testing time was restricted to 30 minutes for each of the three reading tests and for the BCS test and the questionnaire together (see Table 2). The testing time was identical across modes. The initial instruction took about 10 minutes and explained to the test takers how to respond in PBA and CBA, the latter using interactive examples.

Insert Table 2 about here

Randomization check. Students were randomly assigned to either CBA or PBA. To determine whether the random assignment to modes was successful, person variables such as age and gender, score on the reading test for Grade 5, computer usage, BCS score, grade in reading, native language, and self-concept in reading were compared between modes (see Table 3). No statistically significant differences were found between mode groups, supporting the assumption of random equivalence between groups.

Insert Table 3 about here

Data analysis. Data were analyzed in R (R Core Team, 2014) and Mplus (Muthén & Muthén, 1998-2015).

To determine a suitable measurement model, a two-group IRT model was fitted for READ 7 data from both modes using the R package TAM (Kiefer, Robitzsch, & Wu, 2016). Because the test included polytomous items, the Partial Credit Model (PCM; Masters, 1982) and the Generalized Partial Credit Model (GPCM; Muraki, 1992) were used to compare the 1-PL and 2-PL models based on the Akaike Information Criterion (AIC) and Bayesian

Information Criterion (BIC). This analysis was conducted without equality constraints on item parameters.

For analyzing construct equivalence (H_1), the relation to external variables was investigated by comparing the latent correlations of READ 7 (CBA / PBA) with READ 5 (PBA) and o with BCS. Multiple-group IRT models were estimated in Mplus using the Wald test statistic to test the equality of the latent correlations. For this analysis, item difficulties and discriminations were estimated freely.

The comparison of the overall reliability of the paper-based and computer-based tests (H_2) was carried out using the *EAP* reliability obtained from the IRT analysis in TAM (Kiefer et al., 2016). The standard error and confidence intervals of the EAP reliability were estimated using bootstrapping with the SIRT package in R (Robitzsch, 2016) for comparing reliabilities between modes.

The multiple-group IRT models for testing Hypotheses 3.1 to 4.5 were estimated in Mplus (Muthén & Muthén, 1998-2015). In order to test Hypotheses 3.1 to 3.3 regarding differences in item discrimination and difficulty for all items $j = 1...J$, new parameters ME_j were introduced into the multiple-group IRT model, representing the differences in the item parameters under investigation between CBA and PBA for all item pairs. To test the structure of mode effects, constraints were imposed onto these new parameters ME_j. In the model for testing Hypothesis 3.1, $ME_{\alpha j}$ represents the difference between the item discriminations α_j for each item pair of PBA and CBA: $ME_{\alpha j} = \alpha_{j,PBA} - \alpha_{j,CBA}$. Hypothesis 3.1 was tested with the constraint $ME_{\alpha j} = 0$, stating that there is no mode effect on item discriminations for all items j. To test Hypotheses 3.2 and 3.3, $ME_{\beta j}$ expresses the difference in item difficulty β_j for each item pair of PBA and CBA: $ME_{\beta j} = \beta_{j,PBA} - \beta_{j,CBA}$. For PCM items, $ME_{\beta j}$ is the difference between each threshold for an item (i.e., for each threshold the same shift is assumed). The expected mode effect on the test level for item difficulties was tested using the

constraint $\sum ME_{\beta j} = 0$, which means that averaged over all items *j* no mode effect on item difficulties exists. Note that on the test level there might be no mode effect apparent if differences between items cancel out each other. For the analysis of Hypothesis 3.3 on the item level, all differences $ME_{\beta j}$ in item difficulties between PBA and CBA for each item pair were tested by investigating whether $ME_{\beta j}$ differed significantly from zero. A significant difference indicated a mode effect for that particular item.

For our specific hypotheses ($H_{4.1}$ to $H_{4.5}$) on the item level, the mode effect was explained by a set of item properties that were supposed to have an effect on the difference between modes: $ME_{\beta j} = \gamma_1 \cdot x_{j1} + \gamma_2 \cdot x_{j2} + \cdots + \gamma_k \cdot x_{jk}$. For each item property k (e.g., specific response format), x_{jk} indicates whether this property exists for item *j,* with $x_{jk} = 1$ if it exists, and $x_{jk} = 0$ otherwise. Given the identification of this decomposition, the weight γ_k indicates how strongly the property contributes to the mode effect (see Buerger et al., 2016). To incorporate the explanation of mode effects into the multiple-group IRT model, the five item properties ($H_{4.1}$ to $H_{4.5}$, see also Table 1), combo box response format, complex multiple choice response format, multiple choice response format, item position (first and second item) and number of screens per text (more than one screen per text) were added as predictors (see Mplus syntax in digital appendix). As a measure of the effect size, the shift on the logit scale was computed, representing the percentage decrease in the probability of a correct response in CBA for a test taker of average ability.

For all omitted items, not-reached items, and items missing by design, missing responses were ignored in the scaling, meaning that they were treated as not-administered items. This procedure is in line with the scaling of competence data in NEPS (see Pohl & Carstensen, 2012).

Results

Measurement Model and Construct Equivalence

The comparison of the PCM with the GPCM yielded a significant chi-square test (χ^2 = 920.98, df = 43, p <.01), with better model fit for the GPCM. The AIC (39034.21) and BIC (39646.11) of the GPCM were smaller than the AIC (39531.02) and BIC (39935.58) of the PCM. Thus, mode effects were investigated with regard to item discriminations and difficulties.

Furthermore, we used the GPCM to investigate construct equivalence by comparing latent correlations with external variables between modes. When testing for construct equivalence (H_1) using READ 5 as an external variable, the within-mode correlation was 0.88 (SE = 0.02; READ 5 PBA with READ 7 PBA) and was not significantly different from the between-mode correlation of 0.84 (SE = 0.02; READ 5 PBA and READ 7 CBA). The Wald test for the equality constraint of latent correlations in the multiple-group IRT models was not significant, χ^2 = 1.82, df = 1. Using BCS as external variable, the correlation of 0.65 (SE = 0.05; PBA 7 with BCS) was not significantly different from the correlation of 0.53 (SE = 0.06; CBA 7 with BCS; χ^2 = 2.56, df = 1). This means that Hypothesis 1, which postulated equal latent correlations between the reading test and external variables in both modes, was not falsified; thus, evidence for construct equivalence was provided.

Equivalence of Reliability

No differences in the EAP reliabilities were found between PBA, EAP_{PBA} = 0.82, SE_{PBA} = 0.02, $95\%\ CI_{PBA}$ =[0.80, 0.83], and CBA, EAP_{CBA} = 0.80, SE_{CBA} = 0.02, $95\%\ CI_{CBA}$ =[0.76, 0.82], supporting Hypothesis 2. The reliabilities were not statistically different between the two modes.

Psychometric Equivalence of Item Parameters

To investigate Hypothesis 3.1, which assumed equal discriminations between modes, the differences in item discrimination between CBA and PBA for each item were constrained

to be zero for all 42 items simultaneously. The Wald test of the parameter constraints was not statistically significant, indicating that overall there was no difference in item discriminations between the modes and thus no mode effect. Results can be found in Table 4.

The next hypothesis ($H_{3.2}$), which postulated a mode effect on item difficulties on the test level, was tested using the constraint $\sum ME_j = 0$. The Wald test of the parameter constraints was significant, meaning that there was a mode effect apparent on the test level. Consequently, Hypothesis 3.2 was supported. To test Hypothesis 3.2 and Hypothesis 3.3, corresponding item discriminations were fixed to be equal between modes, supported by Hypothesis 3.1.

It was postulated that mode effects are related to item properties and that, due to different item properties, mode effects differ across items ($H_{3.3}$). Thus, mode effects had to be inspected on the item level to test Hypothesis 3.3 regarding differences in difficulty for each item pair (PBA vs. CBA). The significant result of the overall Wald test statistic indicated a mode effect on the item level (see Table 4).

Insert Table 4 about here

Inspecting each item pair revealed a significant mode effect for 22 (of 42) items (see Table 5), supporting the proposition that the mode effect differs across items.

Insert Table 5 about here

Item Properties

Table 6 shows the results regarding whether item properties significantly shift difficulty between the modes. The combo box response format on the compter was expected to affect item difficulty ($H_{4.1}$). Supporting this hypothesis, the analysis showed that items with a combo box response format were systematically more difficult on the computer than the same items administered as assignment tasks on paper. This mode effect was found regardless of whether one had to scroll within the list of entries in the combo box. The position of the correct answer was also shown to be irrelevant. The overall decrease in the probability of

success was 9.10 percent for all combo box items on the computer, that is, the probability of solving items in the combo box response format was 9.10 percent lower in CBA for test takers with an average ability of zero. In addition, no shift in item difficulties was expected for items with multiple-choice and complex multiple-choice response formats ($H_{4.2}$ and $H_{4.3}$). The analysis showed that item difficulties were not shifted significantly in either response format.

Furthermore, the additional navigation required for items where the text was split between multiple screens on the computer were expected to systematically increase item difficulty ($H_{4.4}$). Greater navigation requirements were also expected for items in the first and second position that were not presented on the same page as the text in the computer version, in contrast to the paper version ($H_{4.5}$). As shown in Table 6, splitting the text onto multiple screens had no effect on item difficulty for items on the computer. However, as expected, the additional navigation required to answer the first and second items in a CBA unit significantly increased the item difficulty. For those items, the decrease in the probability of success was 3.52 percent for test takers with an average ability of zero.

Insert Table 6 about here

Missing Value Analysis

The number of not-reached items was expected to be equal for CBA and PBA ($H_{5.1}$). Contrary to this expectation, the number of not-reached items was higher in CBA. More specifically, 8.64 percent of items were not reached on the computer, while 7.79 percent were not reached on paper. The t-test, however, revealed that this difference did not reach statistical significance, $t = .45$, $df = 82$, $p = .66$, supporting Hypothesis 5.1.

Items were expected to be omitted less often under CBA than PBA ($H_{5.2}$). In line with this assumption, 1.29 percent of items were omitted under PBA as compared to 0.30 percent under CBA. The t-test, $t = -4.89$, $df = 82$, $p < .01$, indicated a statistically significant difference, thus confirming Hypothesis 5.2.

Discussion

One main result of this mode effect study indicated that changing the mode from PBA to CBA did not affect the construct measured by the test. Similarly, the reliability of the test was equally high in both modes. While item discriminations were equivalent across modes, a mode effect on item difficulty was found that varied across items, with higher difficulties on average found for items administered on the computer than on paper.

The analysis that included measurement properties to explain mode differences revealed interesting results. Out of three different response formats, only the CBA response format combo box (drop-down menu) significantly increased item difficulty. We found no systematic effect of either the multiple choice response format or the complex multiple choice response format on mode differences in item difficulty. In addition, investigating the effect of additional navigation requirements on the computer showed that splitting reading texts onto multiple screens had no effect on the differences in item difficulty between modes. However, a difference in item difficulty was found for the first and second items of a unit on the computer when these were presented on a different page than the text. CBA items in the first and second position with additional navigation required to read the text turned out to be more difficult than the corresponding PBA items.

The analysis of mode effects for items with varying properties helped identify items that were most likely to be prone to mode effects. Thus, implementing computer-based tests using the combo box format might be not the best way to transfer assignment tasks onto the computer, especially if scrolling is required for some items in a drop down box. Since item difficulty was also affected in items where the drop down box did not require scrolling at all, future implementations should consider using other ways of implementing assignment tasks on the computer. One possibility is to use drag and drop. Such implementation techniques would have to be evaluated in further mode effect studies.

If the first and second items of a unit are presented on the same double-page as the reading stimulus (as is usually the case in PBA), it is easier to respond to them correctly. It is possible that navigating back to the text stimulus after reading the items to filter relevant information leads to higher cognitive workload (e.g. Noyes, Garland, & Robbins, 2004). Pommerich and Burden (2000) recommended a better form of item presentation, suggesting that only part of the text and one item at a time be presented on the computer instead of the full reading text all at once followed by all items. In contrast to the results for navigation requirements concerning the first and second items of a unit, splitting text onto multiple screens does not seem to increase cognitive workload. The relation between additional navigation requirements on the computer and cognitive workload should be investigated in further experimental studies using different navigation types and response formats.

With respect to the number of missing values, we found differences between modes. The computerization of the NEPS Reading test resulted in a significant lower number of omitted items, although the overall number of omitted items was rather small in both modes. This desired mode effect is completely in line with the CBA implementation of a missing value indicator that reminds the test taker of missing answers at the end of each unit. Based on our findings, we suggest following this approach in further computerizations. Additionally, further studies using the missing value indicator should also consider investigating the potential relationship between the (reduced) number of missing values and a mode effect on item difficulty. Specifically, the assumption that missing value indicators pressure test takers to give (wrong) answers (e.g., by random guessing) should be tested. Concerning the number of not-reached items, we found no statistically significant difference between CBA and PBA. This strengthened the assumption that test takers do not actually take advantage of their ability to respond to items in any order, even after completing other units, in paper-based tests (see Kroehne et al., 2016).

The results of our mode effect study raise the question of how to deal with mode effects when comparing the results of paper-based and computer-based tests. If there is sufficient evidence for equivalence in the measured construct, mode-specific item parameters could be used for at least some items, taking into account the change in difficulty. In the case of a heterogeneous mode effect that varies across items, each item with a mode effect gets a specific item parameter, calculated by the shift in difficulty. If the mode effect is homogeneous and affects all items in a similar manner, mode-specific item parameters can be simplified to one mode-specific shift parameter that can be applied to all item difficulty parameters. Estimating mode-specific item parameters -- using random equivalent groups, for instance – makes it possible to handle shifts in item difficulty.

In our study, no mode effect was found on the item level for a considerable proportion of items. In longitudinal studies, which aim to measure change over time, those invariant items could be used to link tests between modes and between time points in order to create a common ability metric that does not require mode-specific item parameters. All other items with mode effects can be retained in the test with mode-specific item parameters, thus accounting for the differences between modes. If the mode effect is related to item properties, which is a major finding of our study, single items or items grouped by an indicator (e.g. “combo box”) might require a mode-specific component. Note, however, that a non-significant mode effect on the item level might also simply be the result of insufficient statistical power. Further research is needed with regard to what size of mode effects are of practical relevance in a given assessment.

A limitation of this research is that the investigation of construct equivalence was limited to available data from a between-subjects design. To more directly investigate the equivalence between CBA and PBA with respect to inter-individual differences, a within-group design would have provided a cross-mode correlation as a criterion for construct equivalence as well as higher statistical power.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington: AERA, APA, NCME.

American Psychological Association, Committee on Professional Standards (COPS) and Committee on Psychological Tests and Assessements (CPTA). (1986). *Guidelines for computer-based tests and interpretations*. Washington: American Psychological Association, Inc.

Artelt, C., Weinert, S., & Carstensen, C. H. (2013). Assessing competencies across the lifespan within the German National Educational Panel Study (NEPS) – Editorial. *Journal for educational research online, 5* (2), 5–14. Retrieved from http://www.pedocs.de/volltexte/2013/8422/pdf/JERO_2013_2_Artelt_Weinert_Carstensen_Editorial_Assessing_competencies.pdf

Association of Test Publishers (ATP). (2000). *Guidelines for computer-based testing*. Washington DC: Association of Test Publishers.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. *The Journal of Technology, Learning, and Assessment, 6* (9).

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.) (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). [Special Issue] *Zeitschrift für Erziehungswissenschaft, 14*.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research, 31* (1), 51–60.

Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The Transition to Computer-Based Testing in Large-Scale Assessments: Investigating (Partial) Measurement Invariance between Modes. *Psychological Test and Assessment Modeling*, *58* (4), 587-606.

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCall 18* (2), 193-211.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and Practices of Test Score Equating* (ETS Research Rep. No. RR-10-29). Princeton, NJ: ETS.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for educational research online, 5* (2), 50–79.

Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing Individual differences in Basic Computer Skills: Psychometric characteristics of an interactive performance measure. *European Journal of Psychological Assessment, 29*, 263-275.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical Guidelines for Assessing Computerized Adaptive Tests. *Journal of Educational Measurement, 21* (4), 347–360. Retrieved from http://www.jstor.org/stable/1434586

Heerwegh, D., & Loosveldt, G. (2002). An Evaluation of the Effect of Response Formats on Data Quality in Web Surveys. *Social Science Computer Review, 20* (4), 471–484.

Holland, P. W., & Dorans, N. J. (2006). Linking and Equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp.187–220). Westport, CT: Praeger.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practices, 20* (3), 16–25.

International Test Commission (ITC). (2005). *International Guidelines on Computer-Based and Internet Delivered Testing.* Retrieved from https://www.intestcom.org/files/guideline_computer_based_testing.pdf

Kiefer, T., Robitzsch, A., & Wu, M. (2016). *TAM: Test analysis modules*. (R package version 1.15-0).

Kim, D.-H., & Huynh, H. (2008). Computer-Based and Paper-and-Pencil Administration Mode Effects on a Statewide End-of-Course English Test. Educational and Psychological Measurement, 68 (4), 554–570.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.)*. New York: Springer-Verlag.

Kroehne, U., Roelke, H., Kuger, S., Goldhammer, F., & Klieme, E. (2016). *Theoretical Framework for Log-Data in Technology-Based Assessments with Empirical Applications from PISA.* Paper presented at the Annual meeting of the National Council on Measurement in Education (NCME), Washington D.C.

Kroehne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft, 14* (2), 169–186.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Mazzeo, J., & von Davier, M. (2008). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. Retrieved 12/12/2008 from http://www.oecd.org/dataoecd/44/49/41731967.pdf

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *ETS Research Report Series,* (1), 1-30.

Muthén, L.K., & Muthén, B.O. (1998-2015). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement, 22* (1), 71–83.

Noyes, J. M., Garland, K. J., & Robbins, E. L. (2004). Paper-based versus computer-based assessment: Is workload another test mode effect? British Journal of Educational Technology, 13(1), 111 - 113.

OECD. (2013). *The Survey of Adult Skills: Reader's Companion.* Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264204027-en

OECD. (2014). *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition, February 2014): Student Performance in Mathematics, Reading and Science*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264208780-en

OECD. (2015). *Adults, Computers and Problem Solving: What's the Problem?* Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264236844-en

OECD. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264266490-en

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26*. *Psychometrics,* (pp.125–167). New York, NY: Elsevier.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large

Scale State Assessment Program. The Journal of Technology, Learning, and Assessment,

3 (6).

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests (NEPS Working Paper No. 14)*. Bamberg: Otto-Friedrich-Universitaet, Nationales Bildungspanel.

Pommerich, M. (2004). Developing Computerized Versions of Paper-and-Pencil Tests: Mode Effects for Passage-Based Tests. *The Journal of Technology, Learning, and Assessment, 2* (6), 3-44.

Pommerich, M. (2007). The effect of using item parameters calibrated from paper administrations in computer adaptive test administrations. *Journal of Technology, Learning and Assessment, 5* (7).

Pommerich, M., & Burden, T. (2000, April). *From Simulation to Application: Examinees React to Computerized Testing.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

Puhan, G., Boughton, K., & Kim, S. (2007). Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. *The Journal of Technology, Learning, and Assessment, 6* (3). Retrieved from http://www.jtla.org

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org

Robitzsch, A. (2016). *Sirt: Supplementary Item Response Theory Models*. (R package version 1.10-0).

Robitzsch, A., Luedtke, O., Koeller, O., Kroehne, U., Goldhammer, F., & Heine, J.-H. (2016). Herausforederungen bei der Schätzung von Trends in Schulleistungsstudien. *Diagnostica.* Retrieved from https://doi.org/10.1026/0012-1924/a000177

Roelke, H. (2012). The ItemBuilder: A Graphical Authoring System for Complex Item Development. *In World Conference on E-Learning in Corporate, Gouvernmet, Healthcare, and Higher Education*. (pp. 344–353), Association for the Advancement of Computers in Education (AACE).

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-Based Testing and Validity: A Look Back and Into the Future. *Assessment in Education, 10* (3), 279-293.

Schroeders, U., & Wilhelm, O. (2010). Testing Reasoning Ability with Handheld Computers, Notebooks, and Paper and Pencil. *European Journal of Psychological Assessment, 26* (4), 284–292.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of Reading and Listening Comprehension Across Test Media. *Educational and Psychological Measurement, 71* (5), 849–869. Retrieved from http://epm.sagepub.com/content/71/5/849

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), London: Lawrence Erlbaum Associates.

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models. *Journal of Educational and Behavioral Statistics, 30* (4), 443–464.

Vispoel, W. P. (2000). Reviewing and Changing Answers on Computerized Fixed-Item Vocabulary Tests. *Educational and Psychological Measurement, 60* (3), 371–384.

Wang, S. (2004). *Online or paper: does delivery affect results? Administration mode comparability study for Stanford Diagnostic Reading and Mathematics Tests*. USA: Pearson Education, Inc.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K 12 Reading Assessments: A Meta-Analysis of Testing Mode Effects. *Educational and Psychological Measurement, 68* (1), 5-24.

Wang, Y.C., Lee, C.M., Lew-Ting, C.Y., Hsiao, C. K., Chen, D.-R., & Chen, W. J. (2005). Survey of substance use among high school students in Taipei: Web-based questionnaire versus paper-and-pencil questionnaire. *Journal of Adolescent Health, 37,* 289–295.

Tables

Table 1

Subsets of Items and Their Properties

	READ 7
Total number of units / reading texts	7
Total number of items	42
Number of partial credit items (*complex multiple choice* and *combo box* items)	15
Investigated Item Sets	
Set 1: *Multiple choice* items	27
Set 2: *Complex multiple choice* items	9
Set 3: *Combo box* items	6
Set 4: Items *with more than one screen* for the reading text stimulus	30
Set 5: Items at the *first or second position* in a unit	14

Note. Each item belongs to Set 1, Set 2 or Set 3 and additionally could belong to Set 4 or Set 5.

Table 2

Administered Instruments

Instrument	Testing Time	
Reading Test Grade 5 (READ 5)	30 min.	Covariate (PBA)
Reading Test Grade 7 (Form A, READ 7A)	30 min.	Outcome (CBA vs. PBA)
Reading Test Grade 7 (Form B, READ 7B)	30 min.	
Basic Computer Skills (BCS)	15 min.	Covariate (CBA)
Questionnaire	15 min.	Covariate (CBA)

Note. READ 7A and READ 7B were combined to READ 7 for analyses.
PBA = administered as paper-based instrument, CBA = administered as CBA instrument.

Table 3

Randomization Check

Covariate	READ 7		
Mode	PBA	CBA	
	M (SE)	*M* (SE)	*t*
Age	13.6 (0.69)	13.56 (0.69)	-0.68
Paper-Based Reading Test 5	1.49 (0.25)	1.47 (0.26)	-1.12
Basic Computer Skills (BCS)	0.76 (0.40)	0.77 (0.33)	0.78
Computer Use	1.4 (0.35)	1.39 (0.34)	-1.44
Questionnaire Reading – Grade	2.84 (0.99)	2.87 (1.03)	0.44
Questionnaire Reading – Native Language	0.94 (1.33)	0.94 (1.23)	0.29
Questionnaire Reading – Self-Concept Reading	3.60 (0.63)	3.65 (0.58)	1.53
Gender			
Male	48 %	46 %	0.82
Female	52 %	54 %	

Note. ** p< .001.

Table 4

Results of Mplus Analysis of Psychometric Equivalence

Hypothesis	Akaike Information Criterion (AIC)	Bayesian Information Criterion (BIC)	χ^2 (df)
Hypothesis 3.1 (item discriminations are equal)	40741.80	41561.04	39.28 (42)
Hypothesis 3.2 (overall mode effect in item difficulty is different from zero)	40583.83	41413.18	26.87 (1)**
Hypothesis 3.3 (mode effect differs across items)	40611.05	41647.74	186.82 (42)**

Note. ** p< .001.

Table 5

Mode Effects on the Item Level

Item No.	Mode Effect (SE)
1	0.26 (0.18)
2	-0.11 (0.24)
3	-0.28 (0.16)
4	0.08 (0.58)
5	-0.01 (0.16)
6	-0.53 (0.09)**
7	-1.08 (0.29)**
8	-0.89 (0.21)**
9	0.04 (0.14)
10	-0.13 (0.19)
11	-0.54 (0.20)**
12	-0.65 (0.10)**
13	-0.45 (0.18)*
14	-0.66 (0.25)**
15	-0.60 (0.18)**
16	-0.38 (0.01)**
17	-0.29 (0.15)**
18	-0.30 (0.15)
19	-0.42 (0.10)**
20	-1.15 (0.22)**
21	-0.19 (0.17)
22	-0.30 (0.12)*
23	-0.74 (0.18)**
24	-0.75 (0.23)**
25	-0.30 (0.18)
26	-0.82 (0.22)**
27	-0.48 (0.22)*
28	-0.66 (0.31)*
29	-0.70 (0.22)**
30	-0.44 (0.08)**
31	-0.24 (0.18)
32	-0.53 (0.14)**
33	0.08 (0.31)
34	-0.11 (0.14)
35	-0.17 (0.22)
36	-0.30 (0.12)**
37	0.33 (0.24)
38	0.16 (0.15)
39	-0.18 (0.25)
40	-0.03 (0.26)
41	0.23 (0.28)
42	-0.22 (0.14)

Note. * p <.05, ** p< .001.

Table 6

Effect of Properties on Mode Effects

Predictor (Hypothesis)	Mode Effect (SE)	Shift on logit scale (%)
Combo box response format ($H_{4.1}$)	-0.36 (0.07)**	9.10
Complex multiple choice response format ($H_{4.2}$)	-0.07 (0.08)	
Multiple choice response format ($H_{4.3}$)	-0.15 (0.10)	
More than one screen per text ($H_{4.4}$)	-0.05 (0.06)	
First and second item (question) not on text page ($H_{4.5}$)	-0.14 (0.07)*	3.52

Note. * p <.05, ** p< .001.

Figures

Figure 1. Illustration of an item with Combo-Box response format.

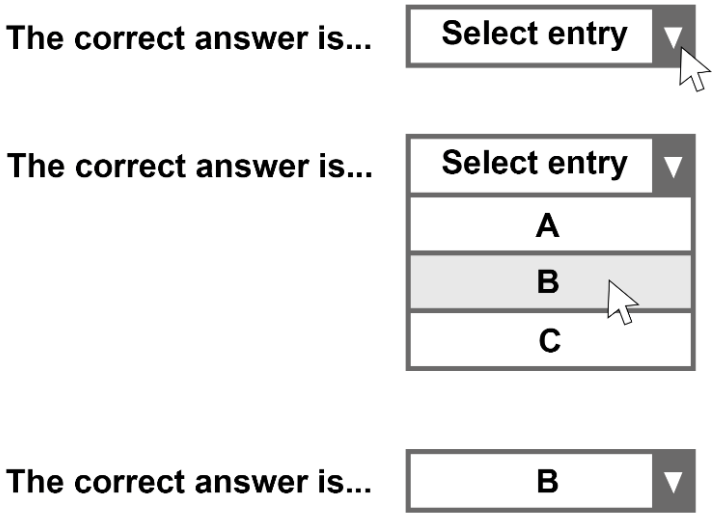
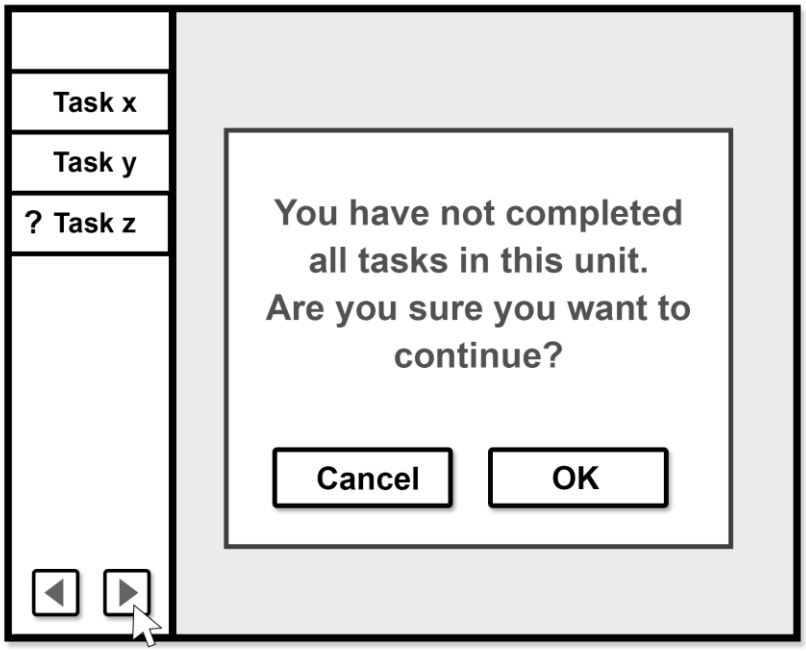


Figure 2. Illustration of the Missing-Value-Indicator.



Footnotes

[1] In the version of the CBA ItemBuilder used in this study, only 5 response options were presented simultaneously in the drop down window. This introduced the additional requirement to scroll down within the *Combo Boxes* to select the sixth item.

Anhang D: Erklärungen zur Promotionsordnung

Eidesstattliche Versicherung

Hiermit bestätigte ich, die vorgelegte, publikationsbasierte Dissertation mit dem Titel *Der Übergang von papierbasiertem zu computerbasiertem Testen in Large-Scale Assessments* selbstständig verfasst und nur die in der Dissertation angegebenen Hilfsmittel in Anspruch genommen zu haben.

Frankfurt am Main, den 20.6.2017 Sarah Suzanne Bürger

Erklärungen über frühere Promotionsversuche

Hiermit erkläre ich, dass keine früheren Promotionsversuche vorliegen.

Frankfurt am Main, den 20.6.2017 Sarah Suzanne Bürger

Erklärung über die Beachtung der Grundsätze der guten wissenschaftlichen Praxis

Ich versichere, die Grundsätze der guten wissenschaftlichen Praxis befolgt zu haben.

Frankfurt am Main, den 20.6.2017 Sarah Suzanne Bürger

Anhang E: Stellungnahme zu den Kriterien einer publikationsbasierten Dissertation

Darlegung der Kriterien *für kumulative Dissertationen im Fachbereich Psychologie und Sportwissenschaften, Goethe Universität Frankfurt* für die vorgelegte, publikationsbasierte Dissertation von Sarah Suzanne Bürger mit dem Titel "Der Übergang von papierbasiertem zu computerbasiertem Testen in Large-Scale Assessments".

(1) Die kumulative Dissertation soll in der Regel 3 Schriften umfassen, die aus den letzten 5 Jahren stammen sollen.

Erklärung:

Folgende Schriften wurden verfasst:

1. *Schrift: Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The Transition to Computer-Based Testing in Large-Scale Assessments: Investigating (Partial) Measurement Invariance between Modes. Psychological Test and Assessment Modeling, 58 (4), 587-606.*
2. *Schrift: Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (submitted, Computers in Human Behavior). Construct Equivalence of PISA Reading Comprehension Measured with Paper-based and Computer-based Assessment.*
3. *Schrift: Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (submitted, Educational and Psychological Measurement). What makes the difference? The Impact of Item Properties on Mode Effects in Reading Assessments.*

(2) Die Schriften sollen im Wesentlichen einem zusammenhängenden Forschungsprogramm entstammen. Die jeweils verfolgten Forschungsfragen sollen sich sinnvoll zueinander in Beziehung setzen lassen.

Erklärung:

Die erste der drei Schriften ist eine theoretisch-methodische Arbeit, die die Analyse von Modus-Effekten (Unterschieden zwischen papierbasiertem und computerbasiertem Assessment) speziell für Large-Scale Assessments schematisch darlegt. Die zweite und dritte Schrift stellen jeweils eine empirische Anwendung einer solchen Modus-Effekt Analyse anhand von Daten aus zwei nationalen Large-Scale Assessments dar (1) Daten aus der Nationalen Begleitforschung aus PISA [PISA NaBe] und 2) Daten des Nationalen Bildungspanels [NEPS].

(3) Der Kandidat oder die Kandidatin soll bei 2 Publikationen Erstautor/Erstautorin sein, bei einer weiteren Publikation kann er/sie Koautor/Koautorin sein. Eine geteilte Erstautorenschaft wird für jeden der Erstautoren anteilig gewichtet (bei 2 Erstautoren eine 1/2 Erstautorenschaft, bei 3 eine 1/3 Erstautorenschaft usw.).

Erklärung:

Die erste und dritte Schrift sind in Erstautorenschaft der Kandidatin verfasst, die zweite Schrift in Koautorenschaft.

(4) Die drei Schriften sollen zur Veröffentlichung zumindest eingereicht sein. Der aktuelle Status ist detailliert darzulegen (Publikationsorgan und Status wie eingereicht, in revision, conditional accept usw.).

Erklärung:

Die erste Schrift ist publiziert in der Zeitschrift „Psychological Test and Assessment Modeling“ (Ausgabe 58 (4), Seiten 587-606.), die zweite Schrift ist eingereicht in der Zeitschrift „Computers in Human Behavior“ (siehe Anhang F) und die dritte Schrift ist eingereicht in der Zeitschrift „Educational and Psychological Measurement“ (siehe Anhang F).

(5) Mindestens 2 der 3 Schriften müssen in guten oder sehr guten, in der Regel englischsprachigen, Zeitschriften mit Peer-Review eingereicht sein.

Erklärung:

Alle der drei oben aufgeführten Zeitschriften sind gute bis sehr gute englischsprachige Zeitschriften mit Peer-Review Verfahren.

(6) Eine der 3 Schriften kann als Publikation in einem einschlägigen Lehrbuch, Enzyklopädieband oder einem anderen für das jeweilige Fach bedeutsamen Publikationsorgan, jeweils mit Peer-Review, eingereicht oder veröffentlicht sein.

Erklärung:

Alle drei Schriften sind in englischsprachigen Zeitschriften mit Peer-Review Verfahren eingereicht.

(7) Die als Dissertation vorgelegte Abhandlung soll über die zusammengestellten Publikationen hinaus einen zusätzlichen Text enthalten, in welchem eine kritische Einordnung der eigenen Publikationen aus einer übergeordneten Perspektive heraus vorgenommen wird. Dieser Text sollte einen Umfang von ca. 30 Seiten haben. Es sollen die Fragestellungen theoretisch entwickelt werden, die empirischen Arbeiten und ihre Ergebnisse so dargestellt werden, dass sie auch ohne Lesen der Einzelarbeiten nachvollziehbar sind und es soll eine Gesamtdiskussion enthalten, die die Fragestellungen beantwortet und den Erkenntnisgewinn der Arbeit herausstellt.

Erklärung:

Die drei Schriften werden von einem zusätzlichen Text gerahmt, der die Fragestellung theoretisch herleitet, die Schriften in deutscher Sprache zusammenfasst und die Arbeiten im Abschluss kritisch diskutiert und reflektiert.

Erklärung über Eigenleistung der Kandidatin

(8) 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. Diese Erklärung ist von Betreuer und/oder Koautoren zu bestätigen.

Erklärung:

1. *Schrift: Die erste Schrift, welche eine methodisch-theoretische Schrift ohne empirische Studie darstellt, wurde von der Kandidatin als Erstautorin verfasst. Sie hat die theoretische Fragestellung mit umfassender Literaturrecherche in Eigenleistung entwickelt. Die theoretisch-methodische Darstellung hat sie daraus selbstständig, unter begleitender Supervision ihres Betreuers sowie der Koautoren abgeleitet und weiterentwickelt sowie kritisch reflektiert.*
2. *Schrift: An der zweiten Schrift war die Kandidatin als Koautorin maßgeblich beteiligt. Hier hat sie wesentliche Teile der Datenanalysen übernommen und selbstständig weitergeführt. Sie hat zudem Teile des Theoriekapitels, der Hypothesendarlegung sowie der Ergebnisdarstellung und Diskussion verfasst.*
3. *Schrift: Die dritte Schrift wurde von der Kandidatin als Erstautorin verfasst. Zur Datengewinnung war sie hier bereits an der Studienplanung sowie Datenerhebung mitbeteiligt. Die zu beantwortenden Fragestellungen hat sie selbstständig entwickelt sowie die statistische Auswertung selbstständig vorgenommen und die*

Schrift in Eigenleistung unter Supervision ihres Betreuers und der Koautoren verfasst.

Prof. Dr. Frank Goldhammer	Sarah Suzanne Bürger
(Betreuer der Dissertation)	(Verfasserin der Dissertation)