Kolumnentitel: UMGANG MIT POSITIONSEFFEKTEN BEIM ADAPTIVEN TESTEN

Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests

Handling of item positions effects in the development of computerized adaptive tests

Andreas Frey[a,b], Raphael Bernhardt[a] und Sebastian Born[a]

[a]Friedrich Schiller University Jena, Germany

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

This is a pre-print version of the article accepted for publication in *Diagnostica.*

This article may not exactly replicate the final version published in the journal. The final version can be found here: https://doi.org/10.1026/0012-1924/a000173

Author Note

Correspondence concerning this article should be addressed to Andreas Frey, Institute of Educational Science, Department of Research Methods in Education, Friedrich Schiller University Jena, Am Planetarium 4, 07743 Jena, Germany, e-mail: andreas.frey@uni-jena.de

Zusammenfassung: Beim computerisierten adaptiven Testen (CAT) werden geschätzte Itemparameter als bekannt und gültig für alle möglichen Darbietungspositionen im Test angesehen. Diese Annahme ist jedoch problematisch, da sich geschätzte Itemparameter empirisch wiederholt als abhängig von der Darbietungsposition erwiesen haben. Die Nichtbeachtung existierender Itempositionseffekte würde zu suboptimaler Itemauswahl und verzerrter Merkmalsschätzung bei CAT führen. Als Lösungsansatz wird ein einfaches Vorgehen zum Umgang mit Itempositionseffekten bei der CAT-Kalibrierung vorgeschlagen. Hierbei werden Item-Response-Theorie-Modelle mit zunehmender Komplexität bezüglich der Modellierung von Positionseffekten geschätzt und das angemessenste Modell aufgrund globaler Modellgeltungskriterien ausgewählt. Das Vorgehen wird an einem empirischen Datensatz aus der Kalibrierung von drei adaptiven Tests (N = 1 632) illustriert. Es zeigten sich Itempositionseffekte, die unterschiedlich differenziert in den einzelnen Tests ausfielen. Durch die Modellierung der Itempositionseffekte wird eine Überschätzung von Varianz und Reliabilität vermieden. Die Nutzung der ermittelten Itempositionseffekte bei nachfolgenden CAT-Anwendungen wird erläutert.

Abstract: In computerized adaptive testing (CAT) item parameter estimates are assumed to be known and valid for all positions items can be presented on within the test. This assumption is problematic since item parameter estimates had been shown to depend upon the position in the test. Neglecting item position effects in CAT-administrations would cause inefficient item selection and biased ability estimation. As a solution, a simple procedure accounting for item position effects is suggested. In this procedure, potential item position effects are identified by fitting and comparing a series of item response theory models with increasing complexity of item position effects. The proposed procedure is illustrated using empirical calibration data of three adaptive tests (N = 1 632). Test specific item position effects were identified. By

accounting for item position effect with an appropriate model, overestimations of variance and reliability were avoided. The implementation of item position effects in operational adaptive tests is explained.

Schlüsselwörter: computerisiertes adaptives Testen, Rasch-Modell, Item-Response-Theorie, Testen, Messen

Keywords: computerized adaptive testing, Rasch model, item response theory, testing, measurement

Kurztitel: Umgang mit Positionseffekten beim adaptiven Testen

Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests

Computerisiertes adaptives Testen (CAT) ist ein spezielles Vorgehen zur Messung individueller Ausprägungen von Personmerkmalen (z. B. Frey, 2012). Dabei wird nach der Beantwortung jedes Items eine vorläufige Schätzung für die individuelle Ausprägung der Testperson im zu messenden Merkmal ermittelt. Anschließend wird aus der Menge der dieser Testperson noch nicht vorgegebenen Items dasjenige Item präsentiert, das für die vorläufige Schätzung ein Optimalitätskriterium (häufig maximale Fisher-Iteminformation) erfüllt. Mit diesem Vorgehen ist ein erheblicher Effizienzzugewinn verbunden, der auf zweierlei Weise genutzt werden kann. Einerseits kann bei gleichbleibender Itemanzahl die Messpräzision im Vergleich zum herkömmlichen Testen mit fester Itemreihenfolge erhöht werden. Wird andererseits die Messpräzision im Vergleich zum herkömmlichen Testen konstant gehalten, kann die Anzahl vorzulegender Items gesenkt werden. Eine Daumenregel besagt, dass bei CAT im Vergleich zum herkömmlichen Testen nur circa die Hälfte der Items ausreicht, um eine vergleichbare Messpräzision zu erreichen (z. B. Segall, 2005). Diesen erheblichen Effizienzvorteil nutzend, findet man computerisierte adaptive Tests heutzutage in vielen Anwendungsbereichen der psychologischen Diagnostik. Neben kommerziell vertriebenen Testverfahren wie zum Beispiel der Intelligenz-Struktur-Batterie (INSBAT; Arendasy et al., 2012) befinden sich darunter auch sehr umfangreiche Testprogramme mit hoher gesellschaftlicher Relevanz. Das aktuell umfangreichste CAT-Programm ist die formative und summative Messung des Erreichens der gemeinsamen Kernstandards für den primären und sekundären Bildungsbereich in den Vereinigten Staaten von Amerika mit mehreren hunderttausend Testungen pro Jahr (z. B. Common Core State Standards Initiative, 2010).

Ein elementarer Bestandteil von CAT ist eine Menge kalibrierter Testitems, der sogenannte Itempool (Thompson & Weiss, 2011). Zur Zusammenstellung des Itempools

werden potentielle Testitems bei einer Kalibrierungsstudie einer großen Personenstichprobe zur Bearbeitung vorgelegt. Die dabei gesammelten Antworten werden zur Schätzung von Itemparametern unter Zuhilfenahme eines Modells der Item-Response-Theorie (IRT; z. B. van der Linden & Hambleton, 1997) verwendet. Die resultierenden Itemparameterschätzungen werden bei der nachfolgenden operationalen Nutzung des adaptiven Tests als gegeben angesehen. Diesem Vorgehen liegt die Annahme zugrunde, dass die geschätzten Itemparameter (a) für unterschiedliche Personen, (b) für unterschiedliche Situationen sowie (c) für unterschiedliche Darbietungspositionen innerhalb des Tests in gleicher Weise gültig sind. Während die Annahme der Invarianz der Itemparameterschätzungen über verschiedene Personengruppen (a) üblicherweise mit Analysen zum Differential Item Functioning (DIF; z. B. Holland & Wainer, 1993) bei der Auswertung der Kalibrierungsstudie geprüft und Variationen der Testergebnisse aufgrund der Durchführung des Tests in unterschiedlichen Situationen (b) durch die Standardisierung von Testbedingungen kontrolliert wird, erfährt die Annahme der Itemparameterinvarianz über Darbietungspositionen (c) bei CAT bislang keine Beachtung.

Das Übersehen der Darbietungsposition kann vermutlich darauf zurückzuführen sein, dass bei herkömmlichen Tests die Itemreihenfolge nicht variiert wird, sodass die Darbietungsposition von Items konstant ist. Selbst wenn Itemparameterschätzungen auf verschiedenen Darbietungspositionen unterschiedlich ausfallen würden, sind keine Verzerrungen der geschätzten individuellen Merkmalsausprägungen zu erwarten, solange die Itemreihenfolge nicht geändert wird. Allenfalls kriteriumsorientierte Testwertinterpretationen (z. B. Herzberg & Frey, 2011) könnten eingeschränkt sein, da die Testwerte eine Mischung der individuellen Merkmalsausprägung und Einflüssen der Itemposition sind. Im Gegensatz zum herkömmlichen Testen mit fester Itemreihenfolge kann bei CAT jedes Item jedoch prinzipiell an jeder Position dargeboten werden. Im Falle eines Effektes der

Darbietungsposition auf die Lösungshäufigkeit (nachfolgend kurz: Itempositionseffekt), und damit auch auf die geschätzten Itemparameter, würden bei CAT für Itemauswahl und Schätzung der individuellen Merkmalsausprägung fast in jedem Fall inkorrekte Itemparameter genutzt. Wurde bei der Kalibrierungsstudie nur eine einzige Testaufgabenanordnung verwendet, dann ist bei Vorliegen von Itempositionseffekten eine Itemparameterschätzung bei CAT nur dann korrekt, wenn das Item an exakt der gleichen Position wie bei der Kalibrierung präsentiert wird. Unterscheidet sich die Darbietungsposition des Items im CAT von der Position in der Kalibrierung, dann sind beispielsweise die bei CAT verwendeten Itemschwierigkeiten je nach Verlauf der Itempositionseffekte entweder zu hoch oder zu niedrig, da die Itempositionseffekte in den Itemschwierigkeiten enthalten sind. Eine ähnlich ungünstige Situation liegt auch dann vor, wenn bei der Kalibrierung unterschiedliche Testzusammenstellungen eingesetzt wurden, die sicherstellen, dass alle Items auf allen Positionen gleich häufig vorgegeben werden, oder wenn die Zuweisung der Items zu Positionen für alle Testpersonen randomisiert erfolgt. In beiden Fällen sind geschätzte Itemparameter bei einer CAT-Anwendung nur an jener Position korrekt, deren Positionseffekt dem Mittelwert aller Positionseffekte entspricht und an allen anderen Positionen nicht.

Liegen Itempositionseffekte vor, ist bei CAT von einer suboptimalen Itemauswahl und von verzerrten Schätzungen der individuellen Merkmalsausprägungen auszugehen. Die suboptimale Itemauswahl resultiert daraus, dass bei vorliegenden Itempositionseffekten während des Testverlaufs systematisch verzerrte vorläufige Merkmalsschätzungen verwendet werden, so dass auch nicht ganz optimale Items ausgewählt werden können. Durch diese Ungenauigkeiten im Testverlauf entsteht unsystematische Fehlervarianz, die sich negativ auf die Präzision der Merkmalsschätzungen auswirkt (z. B. größere Standardfehler). Allein aufgrund von Itemschwierigkeitsschätzungen, die durch Positionseffekte überlagert sind, sind verzerrte Schätzungen der individuellen Merkmalsausprägungen im Sinne von Bias zu

erwarten, die der Summe der Effekte aller Positionen entspricht, auf denen Items bei CAT vorgegeben wurden. Enthält das genutzte IRT-Modell neben Itemschwierigkeiten weitere Itemparameter, ergibt sich eine zunehmend komplexe Gemengelage. Grundsätzlich ist bei Vorliegen von Itempositionseffekten ein Bias bei der Merkmalsschätzung zu erwarten, der in Abhängigkeit der bearbeiteten Items zudem interindividuell unterschiedlich ausfällt. Vor diesem Hintergrund sollte ein herkömmlicher computerisierter adaptiver Test bei Vorliegen von Itempositionseffekten aufgrund der Verletzung der Annahme der Itemparameterinvarianz nicht verwendet werden.

Das skizzierte erhebliche CAT-spezifische Problem wurde bislang noch nicht weiter adressiert und wird von der vorliegenden Arbeit aufgegriffen. Das Ziel besteht dabei in der Vorstellung einer leicht umsetzbaren Möglichkeit, wie bei der Kalibrierung computerisierter adaptiver Tests mit Itempositionseffekte umgegangen werden kann. Konkret wird dabei geprüft, ob Itempositionseffekte vorliegen und wenn ja, von welcher Art diese sind. Das Ergebnis kann in der Folge verwendet werden, um Itempositionseffekte bei der Ausgestaltung des adaptiven Testalgorithmus adäquat zu berücksichtigen. Das vorgeschlagene Vorgehen unterbindet suboptimale Itemselektion und verzerrte Merkmalsschätzung bei CAT aufgrund von Itempositionseffekten. Die Anwendung des Vorgehens wird an einem empirischen Datensatz von $N = 1\,623$ Berufsschülerinnen und Berufsschülern illustriert.

Im weiteren Verlauf werden wir zunächst den Begriff des Itempositionseffekts konkretisieren, die in diesem Bereich bislang vorliegenden empirischen Befunde zusammenfassend darstellen und Ansätze zur psychometrischen Modellierung von Itempositionseffekten beschreiben. Danach fokussieren wir uns auf vier Modelle, die in zielführender Weise zur Identifikation von Itempositionseffekten nutzbar sind und deren Ergebnisse auf einfache Weise bei bestehenden adaptiven Tests genutzt werden können. Hierauf basierend werden die Fragestellungen der Studie spezifiziert, die genutzten Methoden

dargestellt und die Ergebnisse präsentiert. Die Arbeit schließt mit einer Diskussion der Befunde.

Itempositionseffekte

Unter einem Itempositionseffekt verstehen wir die systematische Variation von Itemparametern in Abhängigkeit ihrer Darbietungsposition in einem Test oder Fragebogen. In einer der ersten Arbeiten zu Itempositionseffekten differenziert Mollenkopf (1950) zwischen Itempositioneffekten erster und zweiter Art. Als Itempositionseffekte der ersten Art werden Effekte der Darbietungsposition auf die Itemschwierigkeit und als Itempositionseffekte der zweiten Art Effekte der Darbietungsposition auf die Itemdiskrimination beschrieben.

Itempositionseffekte erster Art.

Itempositionseffekte erster Art waren bei empirischen Studien meist durch einen Abfall der relativen Häufigkeit korrekter Antworten und dem damit einhergehenden Anstieg von Itemschwierigkeiten zum Ende des Tests hin gekennzeichnet (Albano, 2013; Debeer, Buchholz, Hartig & Janssen, 2014; Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Hohensinn et al., 2008; Le, 2007; Meyers, Miller & Way, 2009; Yen, 1980). Zuweilen wurde auch von einem Abfall von Itemschwierigkeiten zum Ende des Tests hin berichtet (Davis & Ferdous, 2005). Die vorliegenden Befunde weisen insgesamt auf ein komplexes Geschehen hin. So berichteten Kingston und Dorans (1984) für die Graduate Record Examination (GRE) sowohl von negativen (Skala „Leseverstehen“) als auch von positiven Effekten (Skala „Analytisches Schreiben“) der Darbietungsposition auf die Itemschwierigkeit. Neben Ergebnissen auf Skalenebene fand Albano (2013) itemspezifische Positionseffekte. Ebenfalls mit GRE-Daten zeigte er, dass zwar auf Skalenebene die Itemschwierigkeiten zum Ende des Tests hin anstiegen, dass dies aber nicht bei allen Items der Skala so war. Es gab auch Items, bei denen spätere Darbietungspositionen mit niedrigeren Itemschwierigkeiten einhergingen. Letztlich liegen Studien vor, bei denen keine oder vernachlässigbare Itempositionseffekte

gefunden wurden (Hohensinn, Kubinger, Reif, Schleicher & Khorramdel, 2011; Li, Cohen & Shen, 2012; Zwick, 1991). Zusammenfassend ist eine relativ heterogene Befundlage zu konstatieren, die darauf hinweist, dass Itempositionseffekte der ersten Art sowohl itemunspezifisch als auch itemspezifisch ausfallen können.

Inhaltliche Erklärungen für das Zustandekommen von Itempositionseffekten liegen bislang nur begrenzt vor. Qian (2014) berichtet für den Test im Schreiben des National Assessment on Educational Progress (NAEP) von substantiellen Zusammenhängen zwischen Itempositionseffekten erster Art mit der Motivation zur Testbearbeitung sowie mit demographischen Variablen wie Geschlecht oder Schulart bei Schülerinnen und Schülern achter und neunter Klassen. In ihrer Zusammenfassung des seinerzeitigen Forschungsstandes zu Itempositionseffekten bezeichneten Leary und Dorans (1985) positive Itempositionseffekte (Abfall der relativen Lösungshäufigkeit zum Ende des Tests hin) als Ermüdungseffekte und negative Itempositionseffekte (Anstieg der relativen Lösungshäufigkeit zum Ende des Tests hin) als Übungseffekte. Diese Bezeichnungen implizieren inhaltliche Begründungen für das Zustandekommen von Itempositionseffekten. Bei genauerer Betrachtung erscheint die Rückführung auf Ermüdung oder Übung zwar intuitiv nachvollziehbar, aber empirisch noch nicht hinreichend untermauert. Im Folgenden werden wir deshalb nicht von Ermüdungs- und Übungseffekten sprechen, sondern den abstrakteren Begriff Itempositionseffekte verwenden.

Itempositionseffekte zweiter Art.

Die vorliegenden empirischen Befunde zu Itempositionseffekten zweiter Art weisen darauf hin, dass Itemdiskriminationen zum Ende von Tests und Fragebögen hin eher ansteigen (Hartig, Hölzel & Moosbrugger, 2007; Knowles, 1988; Le, 2007; Mollenkopf, 1950; Steinberg, 1994; sowie weiterführend Schweizer, Schreiner & Gold, 2009; Schweizer, Troche & Rammsayer, 2011). Jedoch ist auch hier von spezifischen Befundmustern in Abhängigkeit des gemessenen Merkmals beziehungsweise der eingesetzten Skala auszugehen

(Kingston & Dorans, 1984). Inhaltlich könnten Itempositionseffekte der zweiten Art auf eine Gewöhnung an das Testformat oder – bei Persönlichkeitstests – auf eine zunehmende Aktivierung der zu messenden Eigenschaft und/oder des Selbstkonzepts der Testperson zurückzuführen sein, sodass die Antworten zunehmend exakter werden, was sich in höheren Itemdiskriminationen niederschlägt. Allerdings liegen zu inhaltlichen Erklärungen von Itempositionseffekten zweiter Art bislang ebenfalls keine belastbaren empirischen Erkenntnisse vor.

Psychometrische Modellierung.

Eine Möglichkeit zum Umgang mit Itempositionseffekten besteht darin, sie in den genutzten psychometrischen Modellen durch eine geeignete Parametrisierung zu berücksichtigen und damit schätzbar zu machen. Eine Voraussetzung für die Schätzung besteht darin, dass die einzelnen Items auf unterschiedlichen Positionen vorgegeben wurden. Bei Verwendung von nur einer Testzusammenstellung mit fester Itemreihenfolge sind Itemparameter und etwaige Itempositionseffekte indes vollständig konfundiert, sodass letztere nicht geschätzt werden können. Gute Voraussetzungen zur Schätzung von Itempositionseffekten liegen bei Verwendung mehrerer Testzusammenstellungen vor, in denen die Items auf allen Positionen vorgegeben werden. Eine optimale Verteilung von Items auf verschiedene Testzusammenstellungen kann mit balancierten unvollständigen Blockdesigns erreicht werden, die die Itemposition als Blockfaktor enthalten (Frey, Hartig & Rupp, 2009). Die Zuteilung der Testzusammenstellungen zu Testpersonen sollte randomisiert erfolgen.

Zur konkreten Schätzung von Itempositionseffekten können verschiedene psychometrische Modelle eingesetzt werden. Die bislang hierfür vorgeschlagenen Modelle können (a) nach der methodischen Tradition, aus der sie begründet wurden, und (b) nach ihrer Komplexität unterschieden werden. Modelle aus der Tradition der

Strukturgleichungsmodellierung wurden vor allem für Analysen von Itempositionseffekten im Bereich der Persönlichkeitsdiagnostik verwendet (z. B. Schweizer et al., 2009). Mehr und teilweise auch weiterreichende Arbeiten liegen für den Bereich der Leistungsdiagnostik vor. Die Modelle sind fast ausschließlich der IRT zuzuordnen. Sie reichen von einfachen, mehrschrittigen Modellierungsansätzen (Meyers et al., 2009; Pomplun & Ritchie, 2004) über Anwendungen existierender IRT-Modelle wie dem linear-logistischen Testmodell (LLTM; Hohensinn et al., 2008), dem linear-logistischen Testmodell mit Fehlerterm (LLTM+e; Weirich, Hecht & Böhme, 2014) sowie Multi-Facetten-Rasch-Modellen (Li et al., 2012) bis hin zu generalisierten Ansätzen. Eine einflussreiche generalisierte Rahmenkonzeption zur Modellierung von Itempositionseffekten wurde von Debeer und Janssen (2013) vorgelegt. In dieser können Haupteffekte der Position sowie Item x Position-Interaktionen geschätzt werden. Beide sind sowohl als feste Effekte (für alle Personen gleich) als auch als zufällige Effekte (variierend zwischen Personen) spezifizierbar. Ein Großteil der in den letzten Jahren vorgeschlagenen IRT-basierten Modellierungsansätze lassen sich als mehr oder weniger restringierte Formen der generalisierten Rahmenkonzeption von Debeer und Janssen (2013) formulieren.

Vorgeschlagenes Vorgehen

Nachfolgend werden zwei einfach umzusetzende Schritte beschrieben, die standardmäßig bei der Kalibrierung von computerisierten adaptiven Tests eingesetzt werden können. Im ersten Schritt wird eine Abfolge von vier IRT-Modellen geschätzt, bei denen die Modellierung von Itempositionseffekten zunehmend komplexer ausfällt. Im zweiten Schritt wird aufgrund globaler Modellgeltungsindizes entschieden, welche Komplexität hinsichtlich der Modellierung potentieller Itempositionseffekte notwendig ist und damit das bei der späteren CAT-Anwendung zu nutzende Modell festgelegt.

Bei der Wahl der zu schätzenden Modelle ist zu beachten, dass das Ziel im vorliegenden Fall nicht darin besteht, Itempositionseffekte bei einem einzelnen Datensatz möglichst erschöpfend zu modellieren. Vielmehr ist ein Modell zu finden, mit dem der zentrale Anteil der auf Itempositionseffekte zurückgehenden systematischen Varianz so stabil in Form von Parameterschätzungen abgebildet wird, dass diese Parameterschätzungen später bei der Anwendung des adaptiven Tests dauerhaft als gültig angesehen werden können. Dies ist notwendig, da bei CAT-Anwendungen die Parameterschätzungen aus der Kalibrierung bei Itemauswahl und Merkmalsschätzung verwendet und nicht erneut geschätzt werden. Vor diesem Hintergrund ist eine Balance zwischen Differenziertheit der Modellierung von Itempositionseffekten und Übertragbarkeit von Parameterschätzungen zu finden. Damit das vorgeschlagene Vorgehen möglichst breite Anwendung findet, sollten die resultierenden Parameterschätzungen weiterhin möglichst direkt bei CAT nutzbar sein. Da bei CAT feste Itemparameter (im Gegensatz zu personenspezifischen zufälligen Parametern) zum Einsatz kommen, sind für die Itempositionseffekte ebenfalls feste Effekte zu bestimmen und zudem auf Annahmen über den funktionalen Verlauf von Itempositionseffekten vom ersten zum letzten Item (z. B. linear, quadratisch, kubisch, usw.) zu verzichten. Unter diesen Voraussetzungen ist eine direkte Schnittstelle zu existierenden adaptiven Testsystemen gewährleistet, die wir als unverzichtbar ansehen, damit das vorgeschlagene Vorgehen künftig eine breite Anwendung erfährt. Diese erste Arbeit zu der Thematik beschränkt sich ferner auf IRT-Modelle mit einem Itemparameter und dem Rasch-Modell (Rasch, 1966) als Ausgangsmodell. Dieses wird oft bei operationalen adaptiven Tests eingesetzt und wurde auch bei der Entwicklung der adaptiven Tests verwendet, die in der vorliegenden Studie zur Illustration herangezogen werden.

Im Rahmen des vorgeschlagenen Vorgehens wird das Rasch-Modell als *Modell 1* bezeichnet. Es spezifiziert die Wahrscheinlichkeit, dass Person $j = 1,\ldots,N$ Item i korrekt

beantwortet, als Funktion der individuellen Ausprägung des zu messenden Merkmals θ_j und der Itemschwierigkeit b_i:

$$P(U_{ji} = 1|\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \quad (1)$$

Mit dem *Modell 2* wird das Rasch-Modell um den itemunspezifischen Positionsparameter γ_k erweitert zu

$$P(U_{jik} = 1|\theta_j, b_i, \gamma_k) = \frac{\exp(\theta_j - b_i - \gamma_k)}{1 + \exp(\theta_j - b_i - \gamma_k)}. \quad (2)$$

Der hinzugekommene Parameter γ_k beschreibt den Einfluss der Darbietungsposition k = 1,…,K auf die Wahrscheinlichkeit, Item i zu lösen, im Sinne einer positionsspezifischen Schwierigkeit. Das Modell entspricht einem Multi-Facetten-Rasch-Modell (z. B. Linacre, 1994) und ist äquivalent zu dem Modell in Formel 4 von Debeer und Janssen (2013).

Mit dem *Modell 3* wird das Modell 2 um den Interaktionsterm δ_{kl} erweitert. δ_{kl} kann genutzt werden, um die Möglichkeit einzuräumen, dass die Positionsparameter für l = 1,…,L Gruppen von Items (z. B. Wortanzahl des Items; Antwortmodus, usw.) unterschiedlich ausfallen.

$$P(U_{jikl} = 1|\theta_j, b_i, \gamma_k, \delta_{kl}) = \frac{\exp(\theta_j - b_i - \gamma_k - \delta_{kl})}{1 + \exp(\theta_j - b_i - \gamma_k - \delta_{kl})} \quad (3)$$

Das komplexeste Modell stellt das *Modell 4* dar. Mit diesem wird das Modell 2 dahingehend erweitert, dass itemspezifische Positionseffekte durch die Hinzunahme der Facette ζ_{ik} geschätzt werden können:

$$P(U_{jik} = 1|\theta_j, b_i, \gamma_k, \zeta_{ik}) = \frac{\exp(\theta_j - b_i - \gamma_k - \zeta_{ik})}{1 + \exp(\theta_j - b_i - \gamma_k - \zeta_{ik})} \quad (4)$$

ζ_{ik} repräsentiert die Schwierigkeit von Item i auf Position k. Mit diesem Parameter wird systematische Varianz zur Vorhersage der Lösungswahrscheinlichkeit von Item i modelliert,

die über die Schwierigkeit b_i des Items und den Haupteffekt der Darbietungsposition γ_k hinausgeht.

Ein Vorteil der vorgeschlagenen Modelle liegt in ihrer Flexibilität, da für jede Position (Modell 2), jede Position und jede Kombination von Position mit einer als relevant erachteten weiteren Itemeigenschaft (Modell 3) beziehungsweise jede Position und jedes Item auf jeder Position (Modell 4) separate Parameter geschätzt werden. Damit entfallen die mit einigen alternativen Modellierungsansätzen verbundenen und je nach Datenlage mehr oder weniger zutreffenden Annahmen darüber, ob die Itempositionseffekte einem linearen Verlauf oder einem Polynom höherer Ordnung folgen. Gleichzeitig resultiert aus dieser Flexibilität aber auch die Notwendigkeit großer Stichproben, um die Modelle mit ihrer relativ hohen Parameteranzahl hinreichend präzise schätzen zu können. Große Stichproben sind aber bei CAT-Kalibrierungsstudien üblich, sodass hier für gewöhnlich nicht mit Schätzproblemen zu rechnen ist. Sollten beim Modell 4 jedoch zu wenige Antworten pro Item je Position vorliegen, kann durch eine Zusammenfassung von Einzelpositionen zu Positionsgruppen Schätzbarkeit hergestellt werden, ohne dass die Aussagekraft maßgeblich eingeschränkt wird.

Im zweiten Schritt wird entschieden, welches Maß an Komplexität notwendig ist, um etwaige Itempositionseffekte angemessen abzubilden. Um diese Entscheidung zu begründen, werden die dargestellten Modelle anhand globaler Modellgeltungskriterien miteinander verglichen. Bei den vier Modellen handelt es sich um geschachtelte Modelle. Dies ermöglicht einen inferenzstatistischen Vergleich der Modelle auf Basis des χ^2-Differenzentests. Als χ^2-verteilte Prüfstatistik wird hierbei die Differenz der Deviance (-2·log-Likelihood) zwischen den beiden zu vergleichenden Modellen betrachtet, wobei sich die Freiheitsgrade aus der Differenz der Parameteranzahl ergeben. Da die Teststärke des χ^2-Differenzentests von der Stichprobengröße anhängig ist, empfiehlt es sich für den Modellvergleich zusätzlich noch die Informationskriterien BIC (Bayes Information Criterion; Schwarz, 1978), AIC (Akaike‘s

Information Criterion; Akaike, 1978) und CAIC (Consistent AIC; Bozdogan, 1987) hinzuzuziehen. Bei diesen wird die Modellgüte an der Anzahl geschätzter Modellparameter relativiert, sodass eine bessere Modellpassung nur dann ausgewiesen wird, wenn sie nicht mit zu vielen zusätzlichen Parametern „erkauft“ wurde. Je nach Beschaffenheit der Studie sollten dabei jeweils andere Kriterien im Vordergrund stehen (Rost, 2004): Der BIC ist besonders geeignet für Tests mit vielen Items und kleinen Häufigkeiten der beobachteten Antwortmuster. Der AIC empfiehlt sich vor allem bei Tests mit wenig Items und großen Häufigkeiten der beobachteten Antwortmuster. Bei sehr großen Stichproben sollte anstelle des AIC der CAIC Verwendung finden.

Aus der Entscheidung für eines der vier Modelle resultiert die Art der Berücksichtigung etwaiger Itempositionseffekte bei den späteren Anwendungen des adaptiven Tests. Bei Entscheidung für Modell 1 kann die Annahme der Itemparameterinvarianz über Darbietungspositionen beibehalten werden und eine Berücksichtigung von Itempositionseffekten bei der CAT-Anwendung ist nicht notwendig. Bei Entscheidung für Modell 2 sollte bei der CAT-Anwendung der itemunspezifische Positionseffekt berücksichtigt werden. Dies kann bei Rasch-basierten adaptiven Tests auf technisch einfache Weise realisiert werden, indem bei Itemselektion und Merkmalsschätzung der geschätzte Effekt der zum jeweiligen Testzeitpunkt aktuellen Position γ_k zur Itemschwierigkeit b_i addiert wird. Erweist sich Modell 3 als notwendig, dann sind bei Itemselektion und Merkmalsschätzung γ_k und δ_{kl} beziehungsweise bei Modell 4 γ_k und ζ_{ik} zu b_i zu addieren. Bei diesem Vorgehen kann weiterhin das Rasch-Modell und die zugehörige Informationsfunktion verwendet werden, die genutzten Itemschwierigkeiten werden lediglich bei jedem Schritt durch die Addition der entsprechenden Effekte adjustiert. Die direkte Verwendung der Modelle 2 bis 4 bei CAT ist natürlich auch möglich und mit dem zuvor skizzierten Vorgehen äquivalent. Es führt aber zu einem höheren Programmieraufwand.

Fragestellungen

Die Illustration des vorgeschlagenen Vorgehens zum Umgang mit Itempositionseffekten bei der CAT-Kalibrierung orientiert sich an drei Fragestellungen. Zunächst wird untersucht, ob Itempositionseffekte vorliegen, und wenn ja, um welche Art von Itempositionseffekten es sich handelt. Die ersten beiden Fragestellungen lauten entsprechend:

Fragestellung 1: Lassen sich Itempositionseffekte identifizieren?

Fragestellung 2: Sind etwaige Itempositionseffekte (a) für alle Items identisch, (b) für unterschiedliche Gruppen von Items unterschiedlich oder (c) itemspezifisch?

Wenn Itempositionseffekte vorliegen und diese einen stochastischen Zusammenhang mit dem zu messenden Merkmal aufweisen, dann ist bei Verwendung eines Modells ohne explizite Modellierung von Itempositionseffekten zu erwarten, dass systematische Varianz aufgrund von Itempositionseffekten teilweise der Varianz des zu messenden Merkmals σ_θ^2 zugeschrieben wird. Die resultierende Überschätzung der Merkmalsvarianz sowie die damit einhergehende Überschätzung der Reliabilität können durch die Modelle 2, 3 und 4 vermieden werden. Bislang ist nicht bekannt, wie stark die Überschätzung von Varianz und Reliabilität bei Verwendung eines unterspezifizierten Modells bei einer typischen CAT-Kalibrierung ausfällt. Die dritte Fragestellung greift diesen Aspekt auf:

Fragestellung 3: Wie wirkt sich die Modellierung von Itempositionseffekten auf die Varianz und die Reliabilität aus?

Methode

Das vorgeschlagene Vorgehen wird auf Basis der Daten aus der Kalibrierung von drei adaptiven Tests illustriert. Die Tests wurden im Projekt *Messung allgemeiner Kompetenzen–adaptiv* (MaK-adapt; Ziegler, Frey, Seeber, Balkenhol & Bernhardt, 2016) entwickelt, welches in der vom Bundesministerium für Bildung und Forschung (BMBF) geförderten

Forschungsinitiative *Technology-based Assessment of Skills and Competencies in VET* (ASCOT) angesiedelt war.

Stichprobe

Die Stichprobe umfasst N = 1 632 Berufsschülerinnen und Berufsschüler aus 27 Schulen in Hessen, Niedersachsen und Thüringen. Das durchschnittliche Alter der Testpersonen betrug 21.38 Jahre (SD = 3.03). 46 % waren weiblich, für 87 % war Deutsch die Muttersprache und 62 % hatten einen Schulabschluss der mittleren Reife. Die Teilnahme erfolgte freiwillig. Weitere Angaben zur Stichprobe und zur Durchführung der Untersuchung sind Ziegler et al. (2016) zu entnehmen.

Instrumente

Die im Projekt MaK-adapt entwickelten computerisierten adaptiven Tests erfassen Kompetenzen in den Domänen Lesen, Mathematik und Naturwissenschaft. Das theoretische Verständnis hinsichtlich der mathematischen und der naturwissenschaftlichen Kompetenzen orientiert sich an den entsprechenden theoretischen Rahmenkonzeptionen des Programme for International Student Assessment (PISA). Für die Domäne Lesen diente das von Ziegler, Balkenhol, Keimes und Rexing (2012) beschriebene Verständnis funktionaler Lesekompetenz als theoretische Basis.

Die drei Itempools wurden im Wesentlichen aus freigegebenen Items nationaler und internationaler groß angelegter Vergleichsstudien zusammengestellt (vgl. Ziegler et al., 2016). Der bei der Kalibrierung eingesetzte Itempool umfasste 73 Items zur Messung der Lesekompetenz und jeweils 133 Items zur Messung der mathematischen Kompetenz und der naturwissenschaftlichen Kompetenz. Fast alle Items hatten ein geschlossenes Antwortformat. Bei einigen Items waren kurze Angaben wie eine Zahl oder ein Wort in einem Textfeld einzugeben. Die Vorgabe und Bearbeitung der Items erfolgte webbasiert am Computer. Zur Testadministration wurde eine vom Arbeitsbereich Technology-Based-Assessment am

Deutschen Institut für Internationale Pädagogische Forschung angepasste Software zur Vorgabe von mit dem Multidimensional adaptive Testing Environment (MATE; Kroehne & Frey, 2013) erstellten Tests verwendet.

Im Projekt MaK-adapt wurden auf Basis eindimensionaler Skalierungen mit dem Rasch-Modell bei einem multikriterialen Selektionsprozess defizitäre Items identifiziert. Im Ergebnis wurden 105 Mathematikitems, 94 Naturwissenschaftsitems und 65 Leseitems zur Nutzung in den adaptiven Tests ausgewählt. Diese Items bilden die Grundlage der vorliegenden Studie.

Testdesign

Eine Voraussetzung der psychometrischen Analyse von Itempositionseffekten ist, dass die einzelnen Items auf allen möglichen Positionen vorgegeben werden. Für die statistische Modellierung ist es dabei wünschenswert, dass Items und Darbietungspositionen stochastisch unabhängig voneinander sind. Dies könnte prinzipiell mit einer randomisierten Itemauswahl an jeder Darbietungsposition für jede Person erzielt werden. Allerdings stellt sich die erwünschte Gleichverteilung der Items über die Darbietungspositionen erst asymptotisch bei sehr großer Testpersonenanzahl ein. Mit unvollständigen balancierten Blockdesigns kann die erwünschte Gleichverteilung mit geringerer Testpersonenzahl erreicht werden. Bei der vorliegenden Studie kam zur systematischen Zuweisung der Items an die Testpersonen ein balanciertes unvollständiges Blockdesign mit zwei Ebenen zum Einsatz. Auf Ebene 1 des Designs wurden die drei betrachteten Domänen vollständig permutiert, wodurch sechs Sequenzen entstanden (Tabelle 1). Durch die Permutation wurden etwaige Effekte der Domänenabfolge auf das Antwortverhalten ausbalanciert.

Tabelle 1 hier einfügen

Auf Ebene 2 des Designs wurde je Domäne ein sogenanntes Youden square design (YSD) eingesetzt. YSDs sind spezielle balancierte unvollständige Blockdesigns (z. B. Cochran & Cox, 1957), die fünf Eigenschaften genügen:

1. Alle Items kommen gleich häufig vor.

2. Alle Testzusammenstellungen umfassen gleich viele Items.

3. Jedes Item kommt in jeder Testzusammenstellung höchstens einmal vor.

4. Alle Kombinationen von zwei Items in einer Testzusammenstellung kommen gleich häufig vor.

5. Jedes Item kommt an jeder Position gleich häufig vor.

Bei der Kalibrierungsstudie wurden für die Domäne Lesen ein YSD für 73 Items, 73 Testzusammenstellungen (Spalten) und 9 Positionen (Zeilen) sowie für die Domänen Mathematik und Naturwissenschaft jeweils ein YSD für 133 Items, 133 Testzusammenstellungen und 12 Positionen mit dem kostenfreien Computerprogramm Youden 1.0 (Frey & Annageldyev, 2015) konstruiert. Die Designs sind aufgrund ihrer Größe in den elektronischen Supplementen 1 und 2 dieser Publikation zu finden. Die drei domänenspezifischen YSDs wurden in das Permutationsdesign auf Ebene 1 geschachtelt. Entsprechend umfasste die erste Testzusammenstellung neun Leseitems gemäß der ersten Spalte des YSD für die Domäne Lesen, 12 Mathematikitems gemäß der ersten Spalte des YSD für die Domäne Mathematik und 12 Naturwissenschaftsitems gemäß der ersten Spalte des YSD für die Domäne Naturwissenschaft. Für die weiteren Testzusammenstellungen wurden jeweils die folgenden Spalten der YSDs genutzt. Die Sequenz auf Ebene 1 wurde solange beibehalten, bis alle Spalten des YSDs einer Domäne (Ebene 2) mindestens einmal vorgegeben wurden (also 133 Mal). Bei Erreichen der letzten Spalte eines YSD folgte wieder die erste Spalte. Dieses Vorgehen wurde für alle sechs Sequenzen der Ebene 1 wiederholt. Im Ergebnis resultierten 798 Testzusammenstellungen mit jeweils 33 Items aus allen drei

Domänen, von denen jede Testperson eine zufällig ausgewählte zur Bearbeitung vorgelegt bekam.

Versuchsablauf

Nach einer kurzen standardisierten Instruktion und einigen demographischen Fragen wurden die Testpersonen gebeten, die ihnen zugeteilte Testzusammenstellung zu bearbeiten. Jedes Item musste beantwortet werden, ein Überspringen war nicht möglich. Für die Beantwortung der 33 Kompetenzitems standen 45 Minuten zur Verfügung. Die Testung erfolgte ohne größeren Zeitdruck, da die vorgelegten Items von den meisten Testpersonen in der verfügbaren Zeit bearbeitet werden konnten. Nicht erreichte Items wurden als nicht vorgelegt bewertet.

Statistische Analyse

Zur Beantwortung der Fragestellung 1 wurden zunächst die beobachteten relativen Lösungshäufigkeiten in Abhängigkeit der Darbietungsposition deskriptiv betrachtet. Dazu wurde für jede Position die *mittlere relative Lösungshäufigkeit* (Anzahl korrekter Antworten dividiert durch Anzahl aller Antworten) aller an dieser Position vorgegebenen Items in einem Streudiagramm abgetragen. Danach wurde je Domäne Modell 1 mit Modell 2 hinsichtlich der globalen Modellpassung verglichen, um zu entscheiden inwieweit Itempositionseffekte vorliegen.

Für Domänen, bei denen das Modell 2 eine bessere Modellpassung aufwies als das Modell 1, wurde das Modell 2 mit den Modellen 3 und 4 hinsichtlich der Modellpassung verglichen, um zur Beantwortung der Fragestellung 2 zu klären, um welche Art von Positionseffekten es sich handelt. Der bei Modell 3 genutzte Parameter δ_{kl} wurde zur Analyse von Itempositionseffekten in Abhängigkeit der Wortanzahl des Items (kurz: *Itemlänge*) verwendet. Die Anzahl der Wörter je Item variierte von 17 bis 497 Wörter relativ stark. Es sind unterschiedliche Auswirkungen der variierenden Itemlänge auf die Art der Bearbeitung

denkbar. Einerseits könnten Ermüdungseffekte und/oder nachlassende Motivation durch Items mit viel Text intensiviert werden, was zu höheren Itempositionseffekten für lange Items führen würde. Andererseits könnte es bei dem vorliegenden Testformat, bei dem keine Items übersprungen werden können, bei vorliegender Ermüdung und/oder niedriger Motivation insbesondere bei kurzen Items einladend sein, diese auf schnelle Weise durch Raten zu beantworten, was zu höheren Positionseffekten für kurze Items führen würde. Die Kategorien *kurze Items* und *lange Items* wurden mithilfe eines Median-Splits je Domäne gebildet.

Zu erwähnen ist, dass bei der Anpassung der Modelle 2 bis 4 nicht die 33 Einzelpositionen verwendet wurden, sondern Positionsstufen, die mehrere Einzelpositionen zusammenfassen. Bei *Lesen* wurden neun Positionsstufen (jeweils drei Positionen) und bei *Mathematik* und *Naturwissenschaft* sieben Positionsstufen (jeweils fünf Positionen für die ersten sechs Stufen und drei Positionen für die letzte Stufe) gebildet. Im Mittel ergab sich pro Positionsstufe für Lesen eine mittlere Anzahl von 31 Antworten und für Mathematik und Naturwissenschaft von jeweils 34 Antworten. Bei Betrachtung einzelner Positionen wäre die mittlere Anzahl von Antworten je Item-Positions-Kombination mit 10 (Lesen), 7 (Mathematik) und 8 (Naturwissenschaft) zu klein gewesen, um stabile Schätzungen für die itemspezifischen Positionseffekte δ_{ik} bei Modell 4 zu erhalten.

Zur Beantwortung der dritten Fragestellung wurden latente Varianzen sowie EAP/PV-Reliablitäten (Adams, 2005) verglichen.

Für jede Domäne wurden die vier oben beschriebenen Modelle mit der Software ConQuest 3.0.1 (Adams, Wu, Haldane & Sun, 2012) geschätzt. Zur Modellidentifikation wurden die Mittelwerte der Parameter jeder Facette (also: Item, Position, Item*Itemlänge und Item*Position) auf 0 gesetzt. Bei Modell 3 lagen für Lesen und bei Modell 4 für Lesen und Mathematik bei einzelnen Parametern invariante Antwortvektoren vor (nur korrekte, nur

inkorrekte oder nur fehlende Antworten). Die betreffenden nicht schätzbaren Parameter wurden aus den Modellen entfernt.

Ergebnisse

Mit der ersten Fragestellung wird thematisiert, inwieweit in dem analysierten Datensatz Itempositionseffekte identifiziert werden können. Einen ersten diesbezüglichen Einblick liefert Abbildung 1, die die mittlere relative Lösungshäufigkeit aller an derselben Position vorgegebenen Items für die 33 Positionen zeigt. Aus der Abbildung geht zunächst hervor, dass die Itempositionseffekte domänenspezifisch ausfallen. Während die mittleren relativen Lösungshäufigkeiten in den Domänen Lesen und Naturwissenschaft im Testverlauf relativ gleichmäßig abnehmen, sind für die Domäne Mathematik auf den mittleren Positionen die höchsten mittleren relativen Lösungshäufigkeiten zu verzeichnen, wohingegen niedrige und hohe Darbietungspositionen mit niedrigeren mittleren relativen Lösungshäufigkeiten einhergehen. Die zur Hervorhebung des Zusammenhangs zwischen Position und mittlerer relativer Lösungshäufigkeit in der Abbildung 1 abgetragenen linearen und quadratischen Trends unterstreichen dieses Bild. In den Domänen Lesen und Naturwissenschaft können mit linearen Trends erhebliche Varianzanteile der mittleren relativen Lösungshäufigkeit von 32 % beziehungsweise 56 % durch die Darbietungsposition erklärt werden. Durch die quadratische Modellierung kann bei beiden Domänen keine wesentliche Steigerung der Varianzaufklärung erzielt werden. Bei der Domäne Mathematik wird mit dem quadratischen Trend hingegen eine Varianzaufklärung von 42 % erreicht, wobei die Varianzaufklärung bei linearer Modellierung unter 1 % liegt.

Abbildung 1 hier einfügen

Aufbauend auf diesen deskriptiven Betrachtungen wurde die Frage, ob Itempositionseffekte zu beobachten sind, durch den Vergleich der globalen Modellpassung von Modell 1 und Modell 2 beantwortet. Hierbei weist das Modell 2 für alle drei Domänen eine signifikant bessere Modellpassung auf als das Modell 1 (Tabelle 2). Auch BIC, AIC und CAIC sprechen jeweils für das Modell 2. Im Hinblick auf die Fragestellung 1 ist somit zu konstatieren, dass sich in allen drei Domänen Itempositionseffekte identifizieren lassen.

Tabelle 2 hier einfügen

Mit der Fragestellung 2 wird untersucht, ob die gefundenen Effekte der Darbietungsposition für alle Items identisch sind, für lange Items anders ausfallen als für kurze Items oder generell als itemspezifisch anzusehen sind. Dazu wurden Modell 3 mit itemlängenspezifischen Positionseffekten und Modell 4 mit itemspezifischen Positionseffekten für die drei Domänen einzeln geschätzt und jeweils mit Modell 2 bezüglich der globalen Modellgeltung verglichen.

Wie der Tabelle 2 zu entnehmen ist, weisen in den Domänen Lesen und Mathematik weder das Modell 3 noch das Modell 4 eine signifikant bessere Modellpassung auf als das Modell 2. Ebenso sprechen die Werte für BIC, AIC und CAIC für das Modell 2. Für die Domänen Lesen und Mathematik wird deshalb das sparsamere Modell 2 als endgültiges Modell gewählt.

Eine nicht ganz so eindeutige Befundlage zeigte sich in der Domäne Naturwissenschaft. Obgleich auch hier Modell 4 keine signifikant bessere Modellpassung erzielt als Modell 2, fällt die Modellpassung von Modell 3 signifikant besser aus als für Modell 2 . AIC und CAIC bevorzugen ebenfalls das Modell 3, wogegen der BIC das Modell 2 als besser passend ausweist. Vor dem Hintergrund, dass das Modell 3 mit 107 Parametern nur geringfügig

komplexer ist als das Modell 2 mit 101 Parametern und der χ^2-Differenzentest zum Vergleich von Modell 2 und Modell 3 signifikant ausfiel, wurde Modell 3 als endgültiges Modell für die Domäne Naturwissenschaft gewählt.

Bezüglich der Fragestellung 2 wird demnach festgehalten, dass für die untersuchten Tests in den Domänen Lesen und Mathematik von itemunspezifischen Positionseffekten und für die Domäne Naturwissenschaft von itemlängenspezifischen Positionseffekten auszugehen ist.

Zur Beantwortung der dritten Fragestellung wurde abschließend untersucht, wie sich die Modellierung von Itempositionseffekten auf die Varianz und die Reliabilität der gemessenen Merkmalsausprägungen auswirkt. Hierzu wurden je Domäne Varianz und Reliabilität zwischen Modell 1 und dem endgültigen Modell verglichen (Tabelle 3). Da es sich um eine Kalibrierungsstudie handelt, deren Ziel die präzise Schätzung von Itemparametern ist, fallen die Werte erwartungsgemäß relativ niedrig aus. Im direkten Vergleich zeigt sich, dass beide Statistiken beim Modell 1 (ohne Positionseffekte) erwartungsgemäß höher ausfallen als bei den Modellen mit Itempositionseffekten (Tabelle 3). Die Unterschiede sind als marginal (Lesen) bis klein (Mathematik, Naturwissenschaft) einzustufen. Im Ergebnis lässt sich bezüglich Fragestellung 3 festhalten, dass durch die Modellierung von Itempositionseffekten marginale bis kleine Überschätzungen von Varianz und Reliabilität vermieden werden.

Tabelle 3 hier einfügen

Diskussion

Mit der vorliegenden Arbeit wird ein Routineverfahren vorgestellt, mit dem bei der Kalibrierung computerisierter adaptiver Tests geprüft werden kann, ob Itempositionseffekte vorliegen, und wenn ja, von welcher Art diese sind. Das Vorgehen ist auf übliche

computerisierte adaptive Tests auf Basis des Rasch-Modells abgestimmt, sodass die Resultate direkt bei operationalen CAT-Anwendungen genutzt werden können.

Bei der Illustration des Vorgehens anhand empirischer Kalibrierungsdaten zeigten sich deutliche Itempositionseffekte. Diese fielen nicht nur hinsichtlich ihrer Stärke und Differenziertheit, sondern auch bezüglich ihres funktionalen Verlaufs in den drei betrachteten Domänen unterschiedlich aus. In den Domänen Lesen und Naturwissenschaft wurde ein linearer Abfall der mittleren relativen Lösungshäufigkeit im Verlauf des Tests beobachtet. Eine mögliche, wenn auch in der vorliegenden Studie nicht explizit untersuchte, inhaltliche Erklärung für diesen Befund besteht in zunehmender Ermüdung und/oder abnehmender Motivation zur Testbearbeitung im Verlauf der Testung. Im Gegensatz zur den Domänen Lesen und Naturwissenschaft zeigte sich in der Domäne Mathematik zu Beginn der Testung eine geringe mittlere relative Lösungshäufigkeit, die zur Mitte des Tests hin anstieg, um gegen Ende des Tests wieder einzubrechen. Eine inhaltliche Erklärung des anfänglichen Anstiegs der mittleren relativen Lösungshäufigkeit könnte darin liegen, dass insbesondere Items zur Messung mathematischer Kompetenz (z. B. aufgrund ihres vergleichsweise hohen Grades an Abstraktheit) eine Einarbeitungsphase benötigen. Nach Abschluss dieser Einarbeitungsphase könnten dann Ermüdung und/oder abnehmende Motivation zur Testbearbeitung einen Abfall der mittleren relativen Lösungshäufigkeit zum Ende des Tests hin bewirken. Zur Überprüfung der aufgestellten Vermutungen zu den inhaltlichen Erklärungen für Itempositionseffekte wären weitere Studien wünschenswert, bei denen auch geklärt werden sollte, inwieweit der kurvilineare Verlauf bei der Domäne Mathematik replizierbar ist.

Im vorliegenden Fall erwies sich das komplexeste der vier eingesetzten Modelle nicht als das am besten zur Beschreibung der Datenlage geeignete Modell. Vielmehr fiel die Wahl für die Domänen Lesen und Mathematik auf das Modell mit itemunspezifischen

Positionseffekten (Modell 2). In diesen Domänen ändert sich die Lösungswahrscheinlichkeit für alle Items des Tests in Abhängigkeit der Positionen in vergleichbarer Weise. Für die Domäne Naturwissenschaft zeigte sich ein etwas differenzierteres Muster. Hier fiel am Ende des Tests die Lösungswahrscheinlichkeit für kurze Items tendenziell niedriger aus als für lange Items. Dieses Ergebnis könnte dadurch zu erklären sein, dass die Testpersonen aufgrund von Ermüdung und/oder abnehmender Motivation zur Testbearbeitung vor allem bei kurzen Items zum Raten neigen. Sollte es sich hierbei um einen replizierbaren Befund handeln, könnte er bei der Zusammenstellung von Tests zur Verringerung von Itempositionseffekten genutzt werden, indem am Ende des Tests lange Items ohne die Möglichkeit des Überspringens vorgesehen werden.

Der Vergleich der Varianzen und Reliabilitäten zwischen dem Rasch-Modell (Modell 1) und den Modellen mit Itempositionseffekten zeigte, dass wie erwartet Varianz- und Reliabilitätsüberschätzungen durch eine adäquate Modellierung vorhandener Itempositionseffekte vermieden werden können. Die Effekte sind klein, aber systematisch.

Die vorliegende Studie weist in dreierlei Hinsicht praktische Relevanz auf. Zunächst zeigt sie, dass mit substanziellen Itempositionseffekten nicht nur bei mehrstündigen Testungen zu rechnen ist, sondern auch bei vergleichsweise kurzer Testzeit von nur 45 Minuten. Das Problem der Itempositionseffekte kann also nicht durch eine Verkürzung der Testzeit bei der Kalibrierung aufgelöst werden. Vielmehr ist eine adäquate psychometrische Modellierung notwendig. Hierfür bietet die Studie, zweitens, eine Blaupause für ein Routinevorgehen bei der CAT-Kalibrierung an. Die vorgeschlagenen Modelle sind mit üblicher IRT-Software zu schätzen und lassen sich gut kommunizieren. Gleichzeitig sind sie hinreichend komplex, um auch nichtlineare Verläufe von Itempositionseffekten abbilden zu können. Drittens können die aus der Entscheidung für eines der vier Modelle hervorgehenden Itemparameter und Positionsparameter direkt im Rahmen der späteren CAT-Anwendung

genutzt werden. Die Änderungen bestehender CAT-Systeme werden üblicherweise klein ausfallen, sodass sich die Software-Entwicklungskosten in Grenzen halten werden.

Literaturverzeichnis

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

Adams, R. J., Wu, M. L., Haldane, S. & Sun, X. (2012). ACER ConQuest (Version 3.0.1) [Computer software]. Melbourne: ACER Press.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics, 30*, 9–14.

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement, 50*, 408–426.

Arendasy, M., Hornke, L. F., Sommer, M., Wagner-Menghin, M., Gittler, G., Häusler, J. et al. (2012). *Intelligenz-Struktur-Batterie – INSBAT*. Mödling, Österreich: Schuhfried GmbH.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370.

Cochran, W. G. & Cox, G. M. (1957). *Experimental design*. New York: John Wiley & Sons.

Common Core State Standards Initiative (2010). *Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects*. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.

Davis, J. & Ferdous, A. (2005). *Using item difficulty and item position to measure test fatigue.* Washington, DC: American Institutes for Research.

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

Debeer, D. & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*, 164–185.

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. aktualisierte und überarbeitete Auflage, S. 275–293). Berlin, Heidelberg: Springer.

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

Frey, A., Hartig, J. & Rupp, A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53.

Hartig, J. & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, *54*, 418–431.

Hartig, J., Hölzel, B. & Moosbrugger, H. (2007). A confirmatory analysis of item reliability trends (CAIRT): Differentiating true score and error variance in the analysis of item context effects. *Multivariate Behavioral Research Methods, 42*, 157–183.

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

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L. & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly, 50*, 391–402.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E. & Khorramdel, L. (2011). Analyzing item-position effects due to booklet design within large-scale assessment. *Educational Research and Evaluation, 17*, 497–509.

Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale: Erlbaum.

Kingston, N. M. & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*,147–154.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, *55*, 312-320.

Kroehne, U. & Frey, A. (2013). *Multidimensional adaptive testing environment (MATE). Manual*. Frankfurt am Main: DIPF.

Le, L. T. (2007, July). *Effects of item positions on their difficulty and discrimination: A study in PISA Science data across test language and countries*. Paper presented at the 72nd Annual Meeting of the Psychometric Society, Tokyo.

Leary, L. F. & Dorans N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*, 387–413.

Li, F., Cohen, A. & Shen, L. (2012). Investigating the effect of item position in computer-based tests. *Journal of Educational Measurement, 49*, 362–379.

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

Meyers, J. L., Miller, G. E. & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*, 38–60.

Mollenkopf, W. G. (1950). An experimental study of the effects on item analysis data of changing item placement and test-time limit. *Psychometrika, 15*, 291-315.

Pomplun, M. & Ritchie, T. (2004). An investigation of context effects for item randomization within testlets. *Journal of Educational Computing Research, 30,* 243–254.

Qian, J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Psychological Measurement, 38,* 518–534.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, *19,* 49–57.

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

Schwarz, M. (1978). Estimation the dimensions of a model. *Annals of Statistics, 6,* 461–464.

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

Schweizer, K., Troche, S. J. & Rammsayer, T. (2011). On the special relationship between fluid and general intelligence: New evidence obtained by considering the position effect. *Personality and Individual Differences, 50,* 1249–1254.

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 1, p. 429-438). New York, NY: Academic Press.

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*, *66,* 341–349.

Thompson, N. A. & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16,* 1–9. Zugriff am 20.02.2015. Verfügbar unter http://pareonline.net/getvn.asp?v=16&n=1

van der Linden, W. J. & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Weirich, S., Hecht, M. & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38,* 535-548.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*, 297–311.

Ziegler, B., Balkenhol, A., Keimes, C. & Rexing, V. (2012). Diagnostik „funktionaler Lesekompetenz“. *bwp@ Berufs- und Wirtschaftspädagogik–online, 22*, 1–19. Zugriff am 20.5.2016 http://www.bwpat.de/ausgabe22/ziegler_etal_bwpat22.pdf

Ziegler, B., Frey. A., Seeber, S., Balkenhol, A. & Bernhardt, R. (2016). Adaptive Messung allgemeiner Kompetenzen (MaK-adapt). In K. Beck, M. Landenberger & F. Oser (Hrsg.), *Technologiebasierte Kompetenzmessung in der beruflichen Bildung. Ergebnisse aus der BMBF-Förderinitiative ASCOT* (S. 33–54). Bielefeld: wbv.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, *10* (3), 10–16.

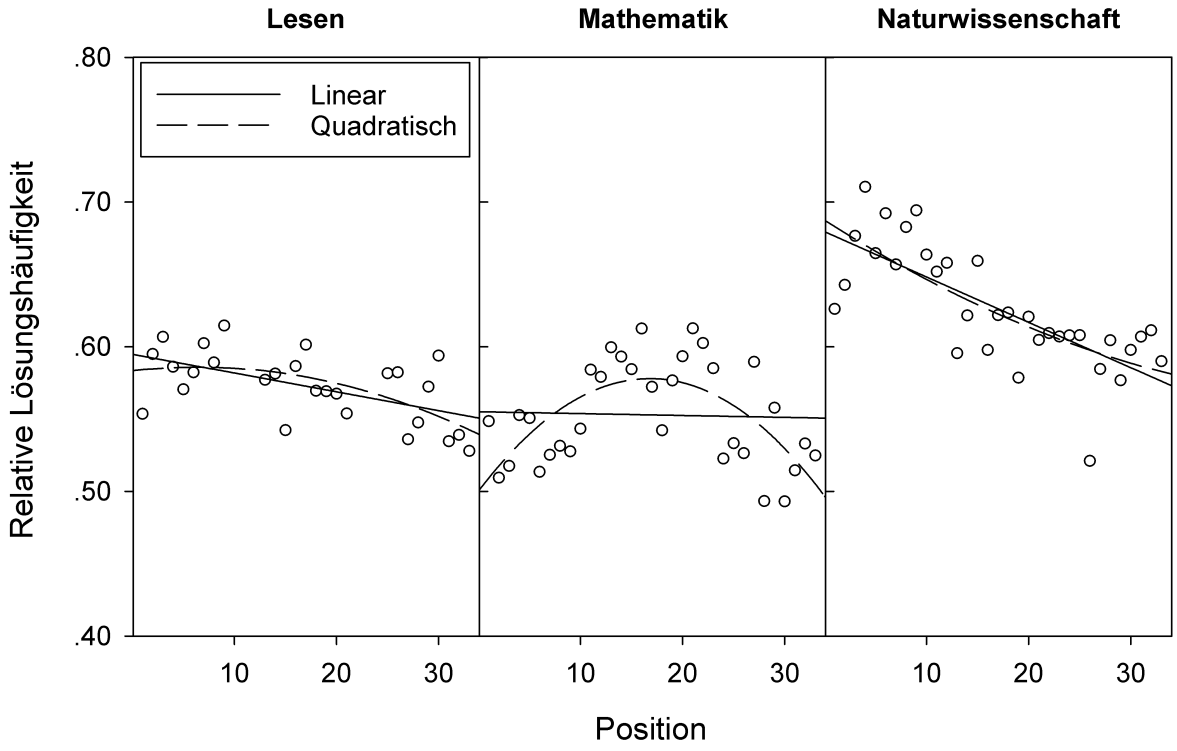

Abbildung 1. Mittlere relative Itemlösungshäufigkeit nach Position und Domäne.

Tabelle 1. Ebene 1 des Testdesigns

	Sequenz					
Block	1	2	3	4	5	6
1	Lesen	Mathe	NaWi	Lesen	Mathe	NaWi
2	Mathe	NaWi	Lesen	NaWi	Lesen	Mathe
3	NaWi	Lesen	Mathe	Mathe	NaWi	Lesen

Anmerkung. NaWi = Naturwissenschaft.

Tabelle 2. Globale Passung für Modelle 1 bis 4 für die Tests in den Domänen Lesen, Mathematik (Mathe) und Naturwissenschaft (NaWi)

Domäne	Modell	Deviance	*p*	BIC	AIC	CAIC
Lesen	1	14 216		14 705	14 348	14 354
	2	14 127	< .001[a]	14 674	14 275	14 282
	3	14 115	.848[b]	14 795	14 299	14 310
	4	13 622	.983[b]	18 416	14 918	15 774
Mathe	1	15 488		16 272	15 700	15 715
	2	15 381	< .001[a]	16 210	15 605	15 622
	3	15 374	.320[b]	16 247	15 610	15 629
	4	14 764	.993[b]	20 823	16 402	18 056
NaWi	1	14 022		14 725	14 212	14 224
	2	13 896	< .001[a]	14 643	14 098	14 111
	3	13 865	< .001[b]	14 656	14 079	14 094
	4	13 351	.644[b]	18 226	14 669	15 564

Anmerkungen. N = 1 632. Deviance = 2·log-Likelihood. p = Irrtumswahrscheinlichkeit χ^2-Differenzentest. BIC: Bayes Information Criterion. AIC: Akaike‘s Information Criterion. CAIC: Consistent AIC. Modell 1: Rasch-Modell. Modell 2: Multi-Facetten-Rasch-Modell mit itemunspezifischen Positionseffekten. Modell 3: Multi-Facetten-Rasch-Modell mit itemunspezifischen Positionseffekten und itemlängenspezifischen Positionseffekten. Modell 4: Multi-Facetten-Rasch-Modell mit itemunspezifischen und itemspezifischen Positionseffekten.

[a] Vergleich mit Modell 1. [b] Vergleich mit Modell 2.

Tabelle 3. Varianz (σ_θ^2) und Reliabilität (Rel)

Domäne	Modell	σ_θ^2	Rel
Lesen	1	0.736	.481
	2	0.732	.479
Mathematik	1	0.953	.560
	2	0.934	.544
Naturwissenschaft	1	0.761	.486
	3	0.722	.458

Anmerkungen. N = 1 632. Modell 1: Rasch-Modell. Modell 2: Multi-Facetten-Rasch-Modell mit itemunspezifischen Positionseffekten. Modell 3: Multi-Facetten-Rasch-Modell mit itemunspezifischen Positionseffekten und itemlängenspezifischen Positionseffekten.