Testbearbeitungsverhalten in Leistungstests: Modellierung von Testabbruch und Leistungsabfall

Dissertation zur Erlangung des Doktorgrades der
Philosophischen Fakultät der
Christian-Albrechts-Universität zu Kiel

vorgelegt von
Dipl. Psych. Marit Kristine List

Kiel
Januar 2018

Erstgutachter:	Prof. Dr. Jens Möller
Zweitgutachter:	Prof. Dr. Gabriel Nagy
Tag der mündlichen Prüfung:	14.05.2018
Durch den Prodekan für Studium und Lehre,	
Prof. Dr. Elmar Eggert, zum Druck genehmigt am:	11.06.2018

Für Großmama.

Danksagung

Bei meiner Arbeit für und rund um meine Dissertation haben mich viele Personen am IPN inhaltlich wie motivational unterstützt, ihnen allen bin ich sehr dankbar. Ich habe in meiner Zeit am IPN viel neues gelernt und mich mit vielen spannenden Forschungsthemen auseinandersetzen können.

Besonderer Dank gilt meinem Doktorvater Gabriel Nagy, für seine Betreuung meiner Arbeit wie auch für seine zahlreichen Ideen und Anregungen. Ich habe in meiner Promotionszeit so viel von ihm lernen können, und bin ihm für all das sehr dankbar.

Von Herzen danken möchte ich Alexander Robitzsch, nicht nur für seine intensive Mitarbeit an meiner zweiten Studie, sondern auch für seine Anregungen zur ersten Studie und ganz besonders für die zahllosen Gespräche über mein Dissertationsthema und darüber hinaus, für seine stete Bereitschaft, methodische Fragen zu beantworten und zu diskutieren. Auch von ihm habe ich dadurch viel gelernt.

Oliver Lüdtke bin ich dankbar für seine Anregungen und Mitgestaltung meiner zweiten Studie und ebenfalls für Rat und Lerngelegenheiten, Karin Guill für ihre Unterstützung im ganzen Prozess meines Promotionsvorhabens.

Und Anni, Dennis, Fabi, Nele *et al.* danke ich für alles Motivieren, Gespräche über Diss und die Welt und das Anfeuern auf den letzten Metern. Viel der Varianz dessen, dass mir meine Promotionszeit in so guter Erinnerung bleibt, wird durch euch aufgeklärt.

An der letzten Position möchte ich auch noch den Personen(variablen) im Hintergrundmodell danken, Johannes und meiner Familie, für eure Geduld und euer Wohlwollen.

Zusammenfassung

In nahezu allen Large-Scale-Assessments (LSAs) bearbeiten nicht alle Testteilnehmenden den Test bis zum Ende mit maximaler Leistung. Testteilnehmende können den Test abbrechen, was sich in der Anzahl der Not-Reached-Items zeigt, oder sie können einen Leistungsabfall zeigen, sodass die Lösungswahrscheinlichkeit am Testende geringer ausfällt. Beide Phänomene stellen eine Veränderung des Testbearbeitungsverhaltens dar. In der Literatur wurden verschiedene Modelle vorgeschlagen, um diese beiden Phänomene im Zusammenhang mit der Testleistung zu untersuchen.

Ziel dieser Dissertation ist, bestehende Modelle zu erweitern, um restriktive Annahmen zu lockern und um die Schätzung differentieller Effekte, das heißt Gruppenunterschiede in der Anzahl der Not-Reached-Items beziehungsweise im Ausmaß des Leistungsabfalls, zu ermöglichen. Im Vordergrund stand die Bewertung der Performanz der Modelle anhand empirischer Datensätze. Im Rahmen der Dissertation wurden zwei Studien durchgeführt, die sich jeweils mit einem der beiden Aspekte der Veränderung des Testbearbeitungsverhaltens (Not-Reached-Items oder Leistungsabfall) beschäftigen.

Die erste Studie der vorliegenden Arbeit beschäftigt sich mit Not-Reached-Items. Not-Reached-Items implizieren einen Testabbruch (z. B. aufgrund von Zeitdruck oder geringer Testbearbeitungsmotivation), der somit eine extreme Form der Veränderung des Testbearbeitungsverhaltens darstellt. In dieser Studie wurde ein Modell zum Testabbruch entwickelt, das eine flexible Modellierung des Zusammenhangs mit der Personenfähigkeit sowie weiterer Kovariaten ermöglicht. Das Modell basiert auf dem Ansatz von Glas und Pimentel (2008), die ein zweidimensionales IRT-Modell vorgeschlagen haben, das neben der Personenfähigkeit eine latente Missingness-Variable, definiert über die Not-Reached-Items, erfasst. Besteht ein signifikanter Zusammenhang zwischen der Personenfähigkeit und der

Missingness-Variable, sind Not-Reached-Items als MNAR zu bewerten (vgl. Little & Rubin, 2002; Schafer & Graham, 2002) und sollten deshalb bei der Parameterschätzung berücksichtigt werden.

Das in der Studie 1 entwickelte *Mixture Discrete (Item) Sequence Event Model* (MDSEM) dient zur Untersuchung von Prädiktoren des Testabbruchs. Das MDSEM hat wie das Modell von Glas und Pimentel (2008) Ähnlichkeiten mit einer Survivalanalyse, in der die Wahrscheinlichkeit, an einer bestimmten Itemposition den Test abzubrechen, in Abhängigkeit von Kovariaten modelliert wird. Die Survivalanalyse bietet ein anschauliches Konzept für die Interpretation der Ergebnisse. Die in der Survivalanalyse üblichen grafischen Darstellungen (z. B. die Survivalfunktion) ermöglichen, auch komplexe Beziehungen in einfacher Weise zu veranschaulichen, was die Interpretation der Ergebnisse in der praktischen Anwendung erleichtert.

Das Modell wurde in zwei empirischen Datensätzen und im Rahmen einer Simulationsstudie mit dem Modell von Glas und Pimentel (2008) und einem Standard-IRT-Modell verglichen und erwies sich den beiden anderen Modellen gegenüber als überlegen. In beiden Stichproben impliziert das MDSEM, dass Personen mit geringerer Mathematikfähigkeit einen späteren Zeitpunkt des Testabbruchs zeigen. Wird das MDSEM um die Gruppenzugehörigkeit als Kovariate erweitert, finden sich in beiden empirischen Datensätzen differentielle Effekte: Unter Kontrolle der Personenfähigkeit zeigt sich, dass die jeweils im Mittel leistungsschwächere Gruppe von Testteilnehmenden höhere Wahrscheinlichkeiten für einen Testabbruch hat. Die Wahrscheinlichkeiten für den Testabbruch sind für Personen mit mittleren und höheren Personenfähigkeiten dabei noch einmal stärker ausgeprägt, es zeigen sich also nichtlineare Effekte für den Testabbruch. Die erste Studie der Dissertation präsentiert mit dem MDSEM eine Erweiterung des Modells von Glas und Pimentel (2008), die die Modellierung von flexiblen Zusammenhängen zwischen Testabbruch und Fähigkeit ermöglicht.

In der zweiten Studie der vorliegenden Arbeit wurden drei Mischverteilungsmodelle von Bolt, Cohen und Wollack (2002), Jin und Wang (2014) und Yamamoto (1995) zur Modellierung von Leistungsabfall miteinander verglichen. Allen drei Modellen liegt die Annahme zugrunde, dass die Stichprobe der Testteilnehmenden aus Personen mit und ohne Leistungsabfall besteht. Ziel der Modelle ist es, die Stichprobe in latente Klassen zu zerlegen, innerhalb derer alle Personen dasselbe Ausmaß von Leistungsabfall zeigen.

Die drei Mischverteilungsmodelle wurden zu Mehrgruppenmodellen erweitert, die ermöglichen, Gruppenunterschiede im Ausmaß des Leistungsabfalls zu schätzen. Die Gruppen können sich dabei einerseits in der Stärke des Leistungsabfalls und andererseits in den Größen der latenten Klassen unterscheiden – also zum einen im Ausmaß der Veränderung des Testbearbeitungsverhaltens und zum anderen in der Anzahl der Personen mit einem Leistungsabfall. Die resultierenden Mehrgruppenmodelle wurden auf einen Mathematiktest angewandt, um Schulformunterschiede im Ausmaß des Leistungsabfalls zu erfassen und um Ähnlichkeiten und Unterschiede in den Implikationen der drei Modelle zu untersuchen.

In allen drei Modellen wird der Anteil der Personen mit Leistungsabfall in den nichtgymnasialen Schulformen höher eingeschätzt, jedoch zeigen sich keine bedeutsamen Gruppenunterschiede im Ausmaß des Leistungsabfalls. Die Modelle von Jin und Wang (2014) und Yamamoto (1995) implizierten, dass Personen ohne Leistungsabfall im Mittel eine höhere Personenfähigkeit haben und dass die Personenfähigkeit umso geringer ist, je früher im Test der Leistungsabfall einsetzt. Das Modell von Bolt et al. (2002) indiziert hingegen keinen Zusammenhang zwischen Personenfähigkeit und Leistungsabfall. Wenn für den Leistungsabfall kontrolliert wird, zeigt sich gegenüber einem Standard-IRT-Modell eine Reduzierung im Schulformunterschied in der Personenfähigkeit.

In beiden Studien wird die Veränderung des Testbearbeitungsverhalten anhand anschaulicher Konzepte formalisiert und die untersuchten Modelle in einem gemeinsamen Rahmen verortet. In der ersten Studie werden die Modelle als Formen einer Survivalanaly-

se, die den Testabbruch untersucht, beschrieben. In der zweiten Studie wird ein generelles Mischverteilungsmodell für den Leistungsabfall entwickelt. In beiden Studien ermöglicht der jeweilige Rahmen, die Unterschiede in den einzelnen Modellen auf einfache Weise darzustellen und zu kontrastieren, indem beschrieben wird, welche Restriktionen für welches Modell eingeführt werden. Die Analysen dieser Dissertation zeigen exemplarisch, wie die vorgestellten Modelle im Rahmen von Sensitivitätsanalysen eingesetzt werden können, um abzuschätzen, wie sehr die Ergebnisse variieren, wenn bestimmte Annahmen zum Testbearbeitungsverhalten modelliert beziehungsweise ignoriert werden.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einführung ins Thema

1.1 Einleitung

Large-Scale-Assessments (LSAs) sind ein wichtiges Instrument im Bildungsmonitoring. Sie dienen der Erfassung von *Personenfähigkeiten* wie etwa Mathematikfähigkeit oder Leseverständnis, um sowohl deren Verteilung in einer Population als auch ihre Veränderung über die Zeit zu schätzen und um Fähigkeitsunterschiede mit Erklärungsvariablen wie Schulform, Geschlecht oder Migrationshintergrund in Beziehung zu setzen (Baumert, 2016; Kato, 2016; Stanat & Artelt, 2009; vgl. Kultusministerkonferenz, 2016).

Die Personenfähigkeiten werden mithilfe von Leistungstests erfasst. Ein Test ist eine Menge von Aufgaben (Items), die eine Testperson lösen (d. h. richtig beantworten) muss. Die Wahrscheinlichkeit, dass eine Person eine Aufgabe richtig löst, hängt von dem Ausmaß ihrer Personenfähigkeit ab. *Personenfähigkeit* stellt somit eine latente Variable dar, deren Werte aus den Itemantworten erschlossen werden (Bartholomew, Knott & Moustaki, 2011; Rost, 1988).

Üblicherweise werden die Leistungstests in LSAs mittels Modellen der Item-Response-Theorie ausgewertet (IRT; vgl. Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). Basierend auf den Annahmen der IRT wird der Zusammenhang der Wahrscheinlichkeit für eine richtige Itemantwort und der zugrunde liegenden latenten Personenfähigkeit modelliert. Ein Vorteil der IRT besteht darin, dass sie ermöglicht, Personenfähigkeiten aus Testversionen, die zwar eine überlappende aber nicht identische Itemmenge haben, auf einer gemeinsamen Metrik darzustellen (*Linking*, vgl. Kolen & Brennan, 2004), sodass die Werte vergleichbar sind.

Testversionen, die jeweils nur eine Itemteilmenge enthalten, sind sinnvoll, um die Testlänge (und damit Testzeit) möglichst gering zu halten, oder um Personen nur Items vorzulegen, die zu ihrem Fähigkeitsniveau passen, sodass sie weder durch zu schwierige Aufgaben

entmutigt oder frustriert noch durch zu einfache Aufgaben gelangweilt werden (vgl. Lilley, Barker & Britton, 2004; S. L. Wise & DeMars, 2005).

Beispiele für LSAs im Kontext des Bildungsmonitorings sind auf internationaler Ebene das *Programme for International Student Assessment* (PISA; vgl. Organisation for Economic Co-operation and Development [OECD], 2017), die *Trends in International Mathematics and Science Study* (TIMSS; vgl. Mullis, Martin & Loveless, 2016) oder die *Progress in International Reading Literacy Study* (PIRLS; vgl. Mullis, Martin, Foy & Drucker, 2012). LSAs auf nationaler Ebene sind in den USA das *National Assessment of Educational Progress* (NAEP; vgl. National Center for Education Statistics [NCES], 2016) oder in Deutschland das nationale Bildungspanel (*National Educational Panel Study* [NEPS]; vgl. Artelt, Weinert & Carstensen, 2013) oder die Überprüfung der Bildungsstandards (*Ländervergleich/ IQB-Bildungstrend*; vgl. Stanat, Schipolowski, Rjosk, Weirich & Haag, 2017). Bei den hier eingesetzten Tests handelt es sich um *Low-Stakes-Tests*, da die Ergebnisse der Testteilnehmenden nicht zur individuellen Leistungsbeurteilung verwendet werden und somit keine Konsequenzen für die Testteilnehmenden haben.

LSAs werden aber auch als *High-Stakes-Tests* eingesetzt, zum Beispiel im Rahmen von universitären Einstufungs- und Aufnahmetests wie dem *Scholastic Aptitude/ Assessment Test* (SAT; s. The College Board, o. J.), dem *Law School Admission Test* (LSAT; s. Law School Admission Council, o. J.) oder dem *Test of English as a Foreign Language* (TOEFL; s. Educational Testing Service, o. J.). Das Ergebnis dieser Tests ist für die Testteilnehmenden relevant, weshalb davon ausgegangen wird, dass in High-Stakes-Tests andere Testbearbeitungsstrategien angewendet werden als in Low-Stakes-Tests (van Barneveld, 2007; Barry, Horst, Finney, Brown & Kopp, 2010; Baumert & Demmrich, 2001; Sundre, 1999; Wolf & Smith, 1995).

Eine der zentralen Fragestellungen an LSAs im Bildungsmonitoring ist die nach Fähigkeitsunterschieden zwischen Gruppen, etwa für die Kategorien Geschlecht oder Migra-

tionshintergrund (Allemann-Ghionda, 2003; Baumert, 2016; Stanat & Artelt, 2009). In Deutschland und anderen Ländern mit einer Leistungsdifferenzierung zwischen verschiedenen Schulformen ist auch von großem Interesse, wie groß die Fähigkeitsunterschiede zwischen verschiedenen Schulformen ausfallen und wie sich diese über die Zeit entwickeln (z. B. Baumert, Maaz & Trautwein, 2009).

Damit Fähigkeitsunterschiede erfasst werden können, müssen die Tests die zu messende Fähigkeit valide, reliabel und in gleicher Weise in den unterschiedlichen Personengruppen abbilden. Das heißt Unterschiede in den Antworten zu den Items des Tests sollen allein auf Unterschiede in der Fähigkeit zurückgeführt werden können und nicht von weiteren Personenmerkmalen oder dem Testdesign abhängen. In der psychometrischen Forschung zu Leistungstests in LSAs wird untersucht, inwieweit diese Ansprüche erfüllt sind: Zentral ist dabei die Untersuchung von Störeinflüssen, die die Grundannahmen der verwendeten IRT-Modelle zur Fähigkeitsschätzung verletzen (vgl. Yen, 1993). Darunter fallen unter anderem fehlende Itemantworten am Testende (*Not-Reached-Items*) und Positionseffekte, zwei typische Phänomene in LSAs, die in einer Vielzahl von Studien dokumentiert und untersucht wurden (für einen Überblick s. Michaelides, 2010; Wu, 2010).

1.2 Not-Reached-Items

Die fehlenden Werte in LSAs können in designbedingt fehlende Werte und fehlende Itemantworten im Test unterteilt werden. *Designbedingt fehlende Werte* kommen dadurch zustande, dass verschiedene Testversionen mit Itemteilmengen eingesetzt werden. Im Rahmen der für LSAs üblichen Methoden zur Parameterschätzung sind designbedingt fehlende Werte unproblematisch (z. B. Walter & Rost, 2011).

Fehlende Itemantworten im Test lassen sich weiter in ausgelassene Items und Not-Reached-Items unterteilen. *Ausgelassene Items* sind alle fehlenden Antworten von Items

im Test, wenn zu den nachfolgenden Items Antworten vorliegen. *Not-Reached-Items* sind alle fehlenden Antworten am Testende, wenn ab einer bestimmten Itemposition auch keines der nachfolgenden Items mehr bearbeitet wurde. Das erste Not-Reached-Item markiert damit einen *Testabbruch*. Konzeptuell sind ausgelassene Items und Not-Reached-Items verschiedene Typen von fehlenden Werten (Lord, 1974, 1980), die durch unterschiedliche Mechanismen entstehen. Diese Dissertation beschäftigt sich mit dem Testabbruch und konzentriert sich deshalb im Folgenden auf Not-Reached-Items. Eine ausführliche Besprechung von ausgelassenen Items findet sich zum Beispiel bei Köhler (2015).

Als Erklärung für Not-Reached-Items werden vor allem Zeitdruck (*Speededness*, vgl. Evans & Reilly, 1972) und geringe Testbearbeitungsmotivation diskutiert. Da die meisten Tests in LSAs aus organisatorischen Gründen zeitlich beschränkt sind, kann es sein, dass nicht alle Personen genug Zeit haben, um den Test bis zum Ende zu bearbeiten (Speededness-Effekt). Eine geringe oder geringer werdende Testbearbeitungsmotivation kann dazu führen, dass eine Person die Testbearbeitung willentlich abbricht. In High-Stakes-Tests erscheinen vor allem Speededness, in Low-Stakes-Tests motivationale Ursachen für Not-Reached-Items plausibel.

In den vorletzten drei PISA-Zyklen (2006, 2009, 2012) zeigte sich, dass im Mittel knapp über 80 % der Testteilnehmenden keine Not-Reached-Items haben, also den Test bis zum Ende bearbeitet haben (OECD, 2009, 2012, 2014). Dem Kriterium des Educational Testing Service zufolge hat ein Test keine zu strikte Zeitbeschränkung, wenn mindestens 80% der Testteilnehmenden das letzte Item erreichen und alle Testteilnehmenden die ersten 75 % der Items bearbeiten (vgl. Bejar, 1985; Lu & Sireci, 2007; Rindler, 1979). Somit liegen die PISA-Tests im Grenzbereich zur Speededness.

Einige Studien berichten von einem Zusammenhang zwischen Not-Reached-Items und Personenfähigkeit für verschiedene Testdomänen, wobei sich sowohl positive wie auch negative Zusammenhänge finden: Debeer, Janssen und De Boeck (2017) berichten für die ar-

gentinische Stichprobe in PISA 2009 einen negativen Zusammenhang für Leseverständnis, sodass die Anzahl der Not-Reached-Items für Personen mit geringerer Fähigkeit größer ist. Negative Zusammenhänge berichten auch Glas und Pimentel (2008) für kognitive Grundfähigkeiten sowie Pohl, Gräfe und Rose (2014) und Wu (2010) für Mathematikfähigkeit. Dahingegen finden Pohl et al. (2014) für einen Leseverständnistest in einer Stichprobe von Schülerinnen und Schülern der fünften Klasse einen positiven Zusammenhang, das heißt Personen mit höherer Fähigkeit weisen mehr Not-Reached-Items auf. Auch G. P. Nagy (1986) berichtet geringe bis mittlere positive Korrelationen zwischen der Anzahl der Not-Reached-Items und der Personenfähigkeit in verschiedenen Tests zur Mathematikfähigkeit und zum Lese- und Sprachverständnis in einer Stichprobe von Schülerinnen und Schülern der vierten Klasse.

Die unterschiedlichen Befunde deuten an, dass der Zusammenhang zwischen Domänen sowie Altersgruppen variiert und auch zwischen Ländern unterschiedlich ausfallen kann. Dies kann ein Hinweis auf differentielle Effekte sein. Differentielle Effekte liegen vor, wenn bei gleichzeitiger Kontrolle der Fähigkeit die Anzahl der Not-Reached-Items auch mit anderen Personenmerkmalen zusammenhängt. Die Untersuchung solch differentieller Raten von Not-Reached-Items ist zum Beispiel bei der Bewertung von Gruppenunterschieden in der Fähigkeit interessant: Da die Kodierung der Not-Reached-Items einen Einfluss auf die Verteilung der Fähigkeitsscores hat, führen differentielle Raten von Not-Reached-Items zu gruppenspezifischen Veränderungen der Verteilung der Scores, je nachdem, wie diese kodiert werden, was (fälschlicherweise) als Unterschied im Fähigkeitsniveau der Gruppen interpretiert werden kann. Außerdem können Unterschiede im Ausmaß von Not-Reached-Items ein Hinweis darauf sein, dass der Test in verschiedenen Gruppen unterschiedlich bearbeitet wird, etwa weil in einer Gruppe vor allem eine zeitintensive Testbearbeitungsstrategie angewendet wird, die dazu führt, dass diese Gruppe den Test nicht in der vorge-

gebenen Zeit beenden kann. Unter Umständen kann das die Validität des Tests sowie der daraus abgeleiteten Statistiken einschränken.

Differentielle Raten von Not-Reached-Items wurden vor allem für Gruppierungsmerkmale wie Geschlecht und Migrationshintergrund untersucht. Während sich für Geschlecht keine Effekte (Bridgeman, Trapani & Curley, 2004; Koretz, Lewis, Skewes-Cox & Burstein, 1993; Schmitt & Crone, 1991) oder nur geringe Effekte (Bridgeman & Cline, 2004; Schmitt, Dorans, Crone & Maneckshana, 1991) finden oder sich die Effekte nicht konsistent in allen untersuchten Tests zeigen (Lawrence, 1993), gibt es eine Reihe von US-amerikanischen Studien, die Unterschiede zwischen Gruppen verschiedener ethnischer Zugehörigkeiten[1] berichten (Evans & Reilly, 1973; Lawrence, 1993; Schmitt & Bleistein, 1987; Schmitt & Dorans, 1988; Schmitt & Crone, 1991).[2] Köhler, Pohl und Carstensen (2015a) finden nur in der Erwachsenenstichprobe einen differentiellen Effekt für Migrationshintergrund, in zwei Schulstichproben derselben Studie zeigt sich kein Effekt.

Testteilnehmende, für die die Testsprache nicht zugleich ihre Muttersprache ist, zeigen bei gleichzeitiger Kontrolle der Fähigkeit mehr Not-Reached-Items (Sireci, Han & Wells, 2008). Eine mögliche Erklärung für die differentiellen Effekte der Testsprache können Unterschiede in der Lesefähigkeit in Mutter- und Zweitsprache sein (Müller & Stanat, 2006), die dazu führen können, dass Testteilnehmende, deren Muttersprache nicht die Testsprache ist, langsamer in der Aufgabenbearbeitung sind und somit eher unter Zeitdruck geraten.

[1] *Ethnische Zugehörigkeit* (engl. *ethnicity*) wird hier im Sinne Webers als soziale Kategorie verstanden (Solga, 2005, S. 261; vgl. Geier & Zaborowski, 2016). Zur Einordnung der Befunde s. Fußnote 2.

[2] Zu beachten ist hier, dass es zwischen den USA und Deutschland aufgrund der kulturellen und historischen Unterschiede unterschiedliche Forschungstraditionen und damit verbunden unterschiedliche Untersuchungsvariablen gibt (z. B. Cokley, 2007; Geier & Zaborowski, 2016), wodurch die Ergebnisse zu *ethnischer Zugehörigkeit* nur bedingt auf Deutschland übertragbar sind, weil hier eher *Migrationshintergrund* untersucht wird (zur Begriffsbestimmung: z. B. Salentin, 2014; Solga, 2005, Kap. 12). Festzuhalten sei, dass die zitierten Befunde darauf hindeuten, dass sich (so verstandene) ethnische oder herkunftsbedingte Disparitäten, die häufig für Leistung gefunden werden (vgl. Dedering & Holtappels, 2010; Kao & Thompson, 2003; Schneider, Martinez & Owens, 2006; Stanat, Rauch & Segeritz, 2010), auch in den Testbearbeitungsstrategien zeigen. Zu beachten ist, dass sowohl ethnische Zugehörigkeit wie auch Migrationshintergrund dabei nicht als Ursache der Unterschiede gesehen werden können, vielmehr gibt es komplexe Mechanismen, die zu den Gruppenunterschieden führen (z. B. Allemann-Ghionda, 2003; Diefenbach, 2007; Stanat et al., 2010).

Für diese Annahme sprechen die Ergebnisse aus der Studie von Köhler et al. (2015a): Köhler et al. (2015a) haben den Zusammenhang von Lesegeschwindigkeit und Not-Reached-Items in NEPS untersucht und finden für alle drei untersuchten Stichproben (5. und 9. Klasse, Erwachsenenstichprobe), dass unter Kontrolle der Fähigkeit die Anzahl der Not-Reached-Items negativ mit der Lesegeschwindigkeit zusammenhängt.

Ländervergleichende Studien zeigen, dass das Ausmaß von Not-Reached-Items variiert (H. H. Chen, von Davier, Yamamoto & Kong, 2015; Debeer et al., 2017; Rose, von Davier & Xu, 2010). Im Vergleich einer französisch- und englischsprachigen Testversion in Kanada finden Emenogu und Childs (2005) ebenfalls Unterschiede in der Anzahl der Not-Reached-Items für beide Sprachgruppen.

Auch zwischen Schulformen gibt es Hinweise auf Unterschiede in den Raten der Not-Reached-Items. In ihrer Analyse der NEPS-Daten finden Köhler et al. (2015a), dass in der neunten Klasse die Anzahl der Not-Reached-Items an den nicht-gymnasialen Schulformen höher ist. In der Stichprobe der fünften Klassen findet sich jedoch kein Unterschied.

1.3 Positionseffekte

Analysen von Leistungstests zeigen konsistent, dass die Lösungswahrscheinlichkeit einer Aufgabe geringer ist, wenn diese an einer späteren Position im Test platziert ist. Dies wird als *Positionseffekt* bezeichnet (z. B. Albano, 2013; Hecht, 2015; Wu, 2010).[3] Positionseffekte können sowohl als Merkmal der Items als auch als Merkmal der Testteilnehmenden betrachtet werden, wobei itemseitige und personenseitige Effekte gleichzeitig auftreten und miteinander interagieren können (für einen Überblick s. Hecht, 2015; Robitzsch, 2009; Wei-

[3]Theoretisch schließen *Positionseffekte* auch positive Effekte ein, die bedeuten, dass die Lösungswahrscheinlichkeit für ein Item mit späterer Position ansteigt (z. B. Übungseffekte, vgl. Yen, 1993). Positive Positionseffekte werden in dieser Arbeit jedoch nicht behandelt, da in Leistungstests negative Positionseffekte deutlich überwiegen (z. B. Bulut, Quo & Gierl, 2017; Hartig & Buchholz, 2012; Hecht, Weirich, Siegle & Frey, 2015; Hohensinn et al., 2008; Le, 2007; G. Nagy, Lüdtke, Köller & Heine, 2017).

rich, 2015). Als ausschließlich itemseitiges Merkmal ist die Stärke des Positionseffekts für alle Testteilnehmenden identisch, kann aber über die Items variieren (z. B. Meyers, Miller & Way, 2009). Zum Beispiel zeigen sich stärkere Positionseffekte für schwierigere Items (Lee & Jia, 2014) und für Items mit einem offenen Antwortformat (Le, 2007).

Wird der Positionseffekt als personenseitiges Merkmal definiert, bedeutet dies, dass sich Personen in ihrem Ausmaß des Positionseffekts unterscheiden (Debeer & Janssen, 2013; Hartig & Buchholz, 2012). Personenseitige Positionseffekte können als *Abfall in der Testleistung* betrachtet werden: Zu Beginn des Tests bearbeiten die Testteilnehmenden die Items mit voller Leistungsfähigkeit, im Testverlauf sinkt der Bearbeitungsaufwand, was zu einer geringeren Lösungswahrscheinlichkeit des Items führt, als gegeben der Personenfähigkeit zu erwarten wäre (vgl. Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Robitzsch, 2009).

In der vorliegenden Arbeit wird der Positionseffekt als Effekt auf der Personenseite verstanden, der durch einen *individuellen Leistungsabfall* entsteht: Je nach Ausmaß des Leistungsabfalls verringert sich im Verlauf des Tests für eine Testperson die Wahrscheinlichkeit, ein Item richtig zu lösen (vgl. Debeer & Janssen, 2013; Hartig & Buchholz, 2012). Ebenso wie Not-Reached-Items kann ein Leistungsabfall sowohl aufgrund von Speededness als auch aufgrund von mangelnder Testbearbeitungsmotivation auftreten (z. B. Bolt et al., 2002; Mittelhaëuser, Béguin & Sijtsma, 2015a). Als weitere Gründe werden unter anderem Ermüdung (Hohensinn et al., 2008; Yen, 1993) oder geringe Selbstkontrollkapazität (Lindner, Nagy, Ramos Arhuis & Retelsdorf, 2017) diskutiert. Ein Leistungsabfall ist sowohl in High-Stakes- als auch in Low-Stakes-Tests denkbar und tritt vermutlich in den beiden Testbedingungen aus unterschiedlichen Gründen auf.

Ähnlich wie der Zusammenhang mit Not-Reached-Items wurde auch die Beziehung zwischen Leistungsabfall und Personenfähigkeit sowie weiteren Personenmerkmalen wiederholt untersucht. Im Bereich der Low-Stakes-Tests wurden vor allem die PISA-Studien

untersucht. Hier finden sich zum Teil positive Zusammenhänge, das heißt, mit höheren Personenfähigkeit geht ein höherer Leistungsabfall einher (Hartig & Buchholz, 2012); zum Teil aber auch negative Zusammenhänge (Debeer & Janssen, 2013; Debeer, Buchholz, Hartig & Janssen, 2014). Debeer et al. (2014) finden allerdings in ihrer Analyse der Leseverständnistests in PISA 2009 für einen Teil der Länder in der Stichprobe keinen Zusammenhang. Im Bereich der High-Stakes-Tests finden Bolt et al. (2002), Cohen, Wollack, Bolt und Mroch (2002) und Goegebeur, De Boeck, Wollack und Cohen (2008) für universitäre Mathematik-Tests ebenfalls negative Zusammenhänge zwischen Leistungsabfall und Personenfähigkeit. Keinen Zusammenhang zwischen Leistungsabfall und Fähigkeit finden hingegen Hailey, Callahan, Azano und Moon (2012).

Wie die Analysen der PISA-Studien zeigen, variiert der Zusammenhang von Leistungsabfall und Personenfähigkeit zwischen Ländern (Debeer et al., 2014; Hartig & Buchholz, 2012). Jin und Wang (2014) berichten Länderunterschiede im Ausmaß des Leistungsabfalls unter Kontrolle der Personenfähigkeit. Darüber hinaus gibt es Hinweise auf Unterschiede im Ausmaß des Leistungsabfalls zwischen Schulen (G. P. Nagy, 1986) oder Schulformen (G. Nagy, Lüdtke & Köller, 2016). Diese Befunde deuten auf differentielle Effekte auf institutioneller und Länderebene hin.

Für Leistungsabfall wurden vielfach differentielle Effekte für Geschlecht, ethnische Zugehörigkeit[4] und Testsprache untersucht. Es finden sich überwiegend keine Geschlechtereffekte (Bolt et al., 2002; Cohen et al., 2002; S. L. Wise, Pastor & Kong, 2009). Studien zu universitären Einstufungs- und Aufnahmetests finden bei gleichzeitiger Kontrolle der Personenfähigkeit Unterschiede im Leistungsabfall zwischen Gruppen verschiedener ethnischer Zugehörigkeit[5] (Bolt et al., 2002; Yamamoto & Everson, 1995) und zwischen Testteilnehmenden, für die die Testsprache die Muttersprache beziehungsweise eine Zweitsprache ist (Schnipke & Scrams, 1997; Yamamoto & Everson, 1997).

[4] Vgl. Fußnote 1.

[5] Zur Einordnung dieser Ergebnisse vgl. Fußnote 2.

Aufgrund der Leistungsdifferenzierung durch die verschiedenen Sekundarschulformen sind *Schulformvergleiche* in Deutschland eine wichtige Untersuchungsvariable in Bildungsvergleichsstudien (Allmendinger, Ebner & Nikolai, 2010; Baumert et al., 2009; Dedering & Holtappels, 2010). Neben den (zu erwartenden) Leistungsdifferenzen zeigen sich für die Schulformen in Deutschland auch differentielle Leistungszuwächse zugunsten der gymnasialen gegenüber den nicht-gymnasialen Schulformen (Becker, Lüdtke, Trautwein & Baumert, 2006; Guill, Lüdtke & Köller, 2017; G. Nagy, Haag, Lüdtke & Köller, 2017; Retelsdorf, Becker, Köller & Möller, 2012; Schiepe-Tiska et al., 2017; für einen Überblick s. Stanat & Artelt, 2009). In ihren Analysen zur Leistungsentwicklung im deutschen PISA-Längsschnitt 2012/2013 konnten G. Nagy, Lüdtke et al. (2017) zeigen, dass der Leistungsabfall an den nicht-gymnasialen Schulformen stärker ausgeprägt war. Werden Schulformunterschiede im Leistungsabfall nicht berücksichtigt, kann sich dies in vermeintlich differentiellen Zuwächsen in der Fähigkeit zeigen (G. Nagy, Retelsdorf, Goldhammer, Schiepe-Tiska & Lüdtke, 2017).

1.4 Veränderung des Testbearbeitungsverhaltens

Sowohl Not-Reached-Items als auch der Leistungsabfall sind Formen einer *Veränderung des Testbearbeitungsverhaltens*. Beide Phänomene können gegen ein Testbearbeitungsverhalten mit anhaltend *maximaler Leistung* kontrastiert werden, das bedeutet, dass Personen den Test bis zum Ende ohne Leistungseinbußen bearbeiten (vgl. Asseburg, 2011). Eine Veränderung im Testbearbeitungsverhalten kann durch verschiedene Ursachen zustande kommen, zum Beispiel aufgrund von Speededness oder geringer Testbearbeitungsmotivation, und in High-Stakes- wie auch in Low-Stakes-Tests auftreten.

In dieser Arbeit erfolgt die Modellierung der Veränderung des Testbearbeitungsverhalten allein anhand der Itemantworten im Test. Daher sind keine Rückschlüsse auf die

Ursachen für Not-Reached-Items oder den Leistungsabfall möglich; hierzu müssten weitere Informationen herangezogen werden, etwa Selbstberichte zur Testbearbeitungsmotivation (Barry et al., 2010; Horst, 2010; Penk & Richter, 2017; Sessoms & Finney, 2015) oder die Bearbeitungszeit, um Speededness zu untersuchen (van der Linden, 2007a; van der Linden, Scrams & Schnipke, 1999).

Alle in dieser Arbeit vorgestellten Modellierungsansätze basieren auf der Annahme, dass die Testteilnehmenden die Items in der im Testheft vorgegebenen Reihenfolge beantworten. Die Definition von Not-Reached-Items setzt diese Annahme voraus. Die zahlreichen Befunde zu Positionseffekten, die einen monotonen Leistungsabfall implizieren (z. B. Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Weirich, 2015), stützen diese Annahme.[6]

1.5 Bestandteile der Dissertation

1.5.1 Thema der Dissertation

Diese Dissertation befasst sich mit den Phänomenen Not-Reached-Items und Leistungsabfall und diskutiert beide als Formen einer Veränderung des Testbearbeitungsverhaltens. In der Literatur wurden verschiedene Ansätze vorgeschlagen, um das Auftreten von Not-Reached-Items und Leistungsabfall zu modellieren. Solche Modelle erlauben Rückschlüsse auf den Zusammenhang mit der Personenfähigkeit und weiteren Kovariaten. Ferner ist es damit möglich, das Ausmaß der Veränderung des Testbearbeitungsverhaltens zu schätzen.

Im Vordergrund dieser Dissertation stehen die Modellierung der Veränderung des Testbearbeitungsverhaltens und der Zusammenhang mit der Personenfähigkeit. Weiter beschäftigt sich diese Arbeit mit der Analyse von Gruppenunterschieden in der Veränderung des

[6] Erwähnt werden soll allerdings, dass auch Modelle für den Leistungsabfall vorgeschlagen wurden, die auf der Annahme basieren, dass Testteilnehmende unabhängig von der vorgegebenen Reihenfolge zuerst die leichten Items bearbeiten, sodass die schwierigen Items vom Leistungsabfall betroffen sind (Bejar, 1985; Y.-W. Chang, Tsai & Hsu, 2014; J. Chang, Tsai, Su & Lin, 2016; Cao & Stokes, 2008). Diese Modelle sind jedoch weniger verbreitet.

Testbearbeitungsverhaltens bei gleichzeitiger Kontrolle von Fähigkeitsunterschieden. Dazu werden bestehende Modellierungsansätze erweitert und miteinander verglichen.

Für diese Dissertation wurden zwei empirische Studien durchgeführt. Die erste Studie (Kap. 4, S. 53ff.) beschäftigt sich mit der Modellierung der Not-Reached-Items als einer Form der Veränderung des Testbearbeitungsverhaltens. Es wird ein Modell entwickelt, das einen bestehenden Ansatz (Glas & Pimentel, 2008) erweitert, um die Analyse von Gruppenunterschieden zu ermöglichen und um komplexe, nichtlineare Beziehungen zwischen Not-Reached-Items und Kovariaten schätzen zu können.

Die zweite Studie (Kap. 5, S. 99ff.) befasst sich mit dem Leistungsabfall als einer anderen Form der Veränderung des Testbearbeitungsverhaltens: In der Literatur wurden verschiedene Modelle vorgeschlagen, um Heterogenität im Ausmaß des Leistungsabfalls innerhalb einer Stichprobe abzubilden. In der zweiten Studie werden drei bekannte Modelle (Bolt et al., 2002; Jin & Wang, 2014; Yamamoto, 1995) erweitert, um zusätzlich Gruppenunterschiede im Testbearbeitungsverhalten schätzen zu können. Die Modelle werden bezüglich ihrer Performanz und ihrer Implikationen miteinander verglichen.

1.5.2 Kapitelübersicht

Die Arbeit ist in sechs Kapitel unterteilt. Kapitel 1 (S. 1ff.) erläutert die relevanten Konzepte bei der Erfassung von Personenfähigkeiten in LSAs und die Formen der Veränderung des Testbearbeitungsverhaltens, Not-Reached-Items und Leistungsabfall, und gibt einen Überblick über die Bestandteile der Dissertation. Kapitel 2 (S. 15ff.) befasst sich mit den messtheoretischen Grundlagen der Fähigkeitsmessung in LSAs, speziell mit den üblicherweise eingesetzten IRT-Modellen. Außerdem werden in diesem Kapitel die Auswirkungen von Not-Reached-Items und Leistungsabfall auf die Fähigkeitsschätzung beschrieben und bisherige Modelle für Not-Reached-Items und Leistungsabfall vorgestellt. Auf Grundlage der Forschungsbefunde und der bestehenden Modellierungsansätze, die in den vorherigen

beiden Kapiteln erörtert werden, werden in Kapitel 3 (S. 45ff.) die Fragestellungen dieser Dissertation abgeleitet. Kapitel 4 (S. 53ff.) und Kapitel 5 (S. 99ff.) beschreiben die im Rahmen der Dissertation durchgeführten Studien. Die Kapitel zu den Studien sind in englischer Sprache verfasst und (in Teilen) in internationalen Fachzeitschriften veröffentlicht. Im Anschluss an die Studienkapitel folgt in Kapitel 6 (S. 139ff.) die Gesamtdiskussion. Sie beinhaltet neben der Zusammenfassung der Ergebnisse eine kritische Würdigung der beiden Studien vor dem Hintergrund bisheriger Forschungsergebnisse und Implikationen für Leistungstests in LSAs.

Kapitel 2

Ansätze zur Modellierung von Testabbruch und Leistungsabfall

2.1 Einleitung und Übersicht

Im vorangegangenen Kapitel wurden zwei typische Probleme in der Datenerhebung in LSAs dargestellt, Not-Reached-Items und Positionseffekte. Beide Phänomene können als Veränderung des Testbearbeitungsverhaltens betrachtet werden, das heißt Testteilnehmende bearbeiten den Test nicht mit maximaler Leistung bis zum Ende, sondern brechen vorzeitig ab oder zeigen einen Leistungsabfall. Leistungstests in LSAs werden üblicherweise mithilfe von Modellen der Item-Response-Theorie ausgewertet (IRT, vgl. Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). Not-Reached-Items und Leistungsabfall bleiben in diesen Modellen unberücksichtigt.

Im folgenden werden die Grundlagen der IRT-Modelle dargestellt. Anschließend wird die Bedeutung von Not-Reached-Items und Leistungsabfall für die Schätzung der Personenfähigkeit und anderer Modellparameter erläutert und allgemeine Modellierungsansätze für Not-Reached-Items und für den Leistungsabfall dargestellt. Dazu wird zunächst im Abschnitt 2.3 auf Not-Reached-Items und im nachfolgenden Abschnitt 2.4 auf den Leistungsabfall eingegangen.

2.2 Item-Response-Theorie-Modelle für Leistungstests

Die Grundidee der IRT ist, dass die beobachteten Itemantworten auf zugrunde liegende Personenfähigkeiten und Eigenschaften der Items zurückgeführt werden. Die in der vorliegenden Arbeit verwendeten Datensätze bestehen ausschließlich aus dichotomen Items ($0 =$ *falsch gelöst*, $1 =$ *richtig gelöst*), beispielsweise Multiple-Choice-Aufgaben mit genau einer richtigen Lösung, und die Tests erfassen genau eine Personenfähigkeit, etwa die *Ma-*

thematikfähigkeit. Deshalb berücksichtigt die folgende Beschreibung der IRT-Modelle auch nur diesen Itemtyp und eine Fähigkeitsdimension.[7]

IRT-Modelle gehen davon aus, dass die Ausprägung einer Person auf der latenten Fähigkeitsvariable deren Antwortverhalten auf den Items bedingt. Die Beziehung zwischen Itemantworten und Fähigkeitsausprägung wird durch eine logistische Regression von der Wahrscheinlichkeit einer richtigen Itemlösung auf die Personenfähigkeit und Itemeigenschaften geschätzt. Die verschiedenen IRT-Modelle unterscheiden sich darin, welche Itemeigenschaften modelliert werden. Das wohl bekannteste Modell ist das *Rasch-Modell* oder das 1-Parameter-logistische Modell (1PL; Rasch, 1980), in dem angenommen wird, dass sich Items nur in ihrer *Itemschwierigkeit* unterscheiden. Für eine Person p ergibt sich die Wahrscheinlichkeit, ein Item i richtig zu lösen, $P(X_{pi} = 1)$, aus ihrer Ausprägung auf der Fähigkeitsvariable θ_p und der Itemschwierigkeit b_i:

$$P(X_{pi} = 1) = \frac{\exp(\theta_p - b_i)}{1 + \exp(\theta_p - b_i)}. \tag{2.1}$$

Gebräuchlicher ist die Schreibweise in der Logitform. Der Logit η_{pi} ist der natürliche Logarithmus des Wettquotienten, im IRT-Modell für dichotome Items ist dies das Verhältnis von Wahrscheinlichkeit einer richtigen und einer falschen Itemlösung:

$$\eta_{pi} = \operatorname{logit} P(X_{pi} = 1) = \ln \frac{P(X_{pi} = 1)}{1 - P(X_{pi} = 1)}.$$

In Logitform wird Gleichung 2.1 zu:

$$\eta_{pi} = \theta_p - b_i. \tag{2.2}$$

[7] Ausführliche Darstellungen von IRT-Modellen und deren Erweiterungen für mehrstufige Items oder mehrere Fähigkeitsdimensionen finden sich zum Beispiel bei Embretson und Reise (2000); Hambleton und Swaminathan (1985); Lord (1980); Yen und Fitzpatrick (2006).

Die Metrik der Itemschwierigkeiten der Items eines Tests entspricht der Metrik der Personenfähigkeitsvariable. Sind Itemschwierigkeit und Personenfähigkeit identisch, beträgt die Wahrscheinlichkeit, das Item richtig zu lösen, 50 % (s. Gl. 2.2):

$$\frac{\exp(\theta_p - b_i)}{1 + \exp(\theta_p - b_i)} \Leftrightarrow \frac{\exp(0)}{1 + \exp(0)} = 0.5.$$

Leichtere Items werden mit einer Wahrscheinlichkeit größer als 50 %, schwierigere Items mit einer Wahrscheinlichkeit kleiner als 50 % richtig beantwortet.

Die weiteren IRT-Modelle unterscheiden sich vom 1PL darin, dass sie noch weitere Itemeigenschaften definieren. Das 2-Parameter-logistische Modell (2PL; Birnbaum, 1968) nimmt an, dass sich Items auch in ihrer Diskriminationskraft unterscheiden können, was durch den *Diskriminationsparamater* ausgedrückt wird, der neben Itemschwierigkeit die Antwortwahrscheinlichkeit bedingt. Praktisch bedeutet dies, dass sich Items darin unterscheiden können, wie gut sie zwischen Abstufungen auf der Fähigkeitsskala unterscheiden können. Dahingegen wird im 1PL implizit angenommen, dass alle Items einen gleich hohe Diskriminationskraft aufweisen, das heißt denselben Diskriminationsparameter haben. Für das 2PL wird Gleichung 2.2 um den itemspezifischen Diskriminationsparameter a_i erweitert:

$$\eta_{pi} = a_i \cdot (\theta_p - b_i). \tag{2.3}$$

Birnbaum (1968) hat darüber hinaus ein Modell vorgeschlagen, dass eine untere Asymptote der Lösungswahrscheinlichkeit als Ratewahrscheinlichkeit im Modell ergänzt (3-Parameter-logistisches Modell: 3PL), sodass auch bei sehr geringen Personenfähigkeiten die Lösungswahrscheinlichkeit eines Items größer Null ist. Als Erweiterung zum 3PL haben Barton und Lord (1981) das 4-Parameter-logistische Modell (4PL) vorgeschlagen, dass zusätzlich eine obere Asymptote für die Lösungswahrscheinlichkeit annimmt. Das 4PL nimmt

an, dass auch Testteilnehmende mit einer sehr hohen Personenfähigkeit ein Item nicht immer richtig lösen werden, was im Modell dadurch realisiert wird, dass die Lösungswahrscheinlichkeit auch für hohe Werte auf der Fähigkeitsvariable kleiner 1 geschätzt wird. Das 4PL hat erst in den letzten Jahren Beachtung erfahren (s. Magis, 2013); so wurde es zum Beispiel von Culpepper (2017) zur Modellierung von geringer Testbearbeitungsmotivation vorgeschlagen.

Nationale und internationale Bildungsvergleichsstudien verwenden unterschiedliche IRT-Modelle. PISA hat bis 2012 das 1PL verwendet und seit 2015 werden die PISA-Leistungstests mit dem 2PL ausgewertet (OECD, 2017). Im NEPS wird das 1PL verwendet (Pohl & Carstensen, 2012). TIMSS, PIRLS und NAEP verwenden für Multiple-Choice-Items das 3PL, für dichotome Items mit einem offenen Antwortformat das 2PL (Foy, Brossman & Galia, 2012; NCES, 2016; Martin, Mullis & Hooper, 2016). Das 4PL wird bisher nicht in den großen Bildungsvergleichsstudien eingesetzt.

Auch wenn die komplexeren Modelle wie das 3PL und das 4PL plausiblere Annahmen über das Testbearbeitungsverhalten treffen, decken auch sie bei weitem nicht alle Aspekte ab. Die Entscheidung für ein bestimmtes IRT-Modell beruht meist eher auf pragmatischen Gründen. Mit dem Gewinn an Flexibilität (und Plausibilität) in den höher parametrisierten Modellen (2PL, 3PL, 4PL) gehen auch eine größere Komplexität und dadurch bedingt längere Laufzeiten bei der Schätzung der Modellparameter und – werden nicht zusätzliche Restriktionen auf die Parameter eingeführt – die Notwendigkeit größerer Stichproben einher. Ferner ist die Schätzung eines 3PL oder 4PL nicht in allen gängigen Software-Paketen möglich, was ein weiterer pragmatischer Grund für die Anwendung des 1PL oder 2PL ist.

Die IRT-Modelle, die in LSAs zur Erfassung der Personenfähigkeiten eingesetzt werden (vgl. Gl. 2.2 und 2.3), gehen davon aus, dass Testteilnehmende alle Items des Tests mit maximaler Leistung bearbeiten. Es wird also angenommen, dass sich das Testbearbeitungsverhalten im Testverlauf nicht verändert, sodass die Testteilnehmenden weder den

Test abbrechen (und dadurch fehlende Werte durch die Not-Reached-Items haben) noch einen Leistungsabfall zeigen.

Die in LSAs üblichen Schätzverfahren können auch mit unvollständigen Daten umgehen, sodass fehlende Werte für die Parameterschätzung kein Problem darstellen, sofern das Fehlen unter Berücksichtigung der beobachteten Werte unsystematisch ist (Mislevy & Wu, 1996). Wie oben dargestellt wurde, hängt das Fehlen von Itemantworten jedoch in vielen Fällen von der Ausprägung auf der Fähigkeitsvariable ab, und somit ist diese Annahme verletzt.

In den IRT-Modellen (vgl. Gl. 2.2 und 2.3) gibt es keine weiteren systematischen Merkmale außer der Personenfähigkeit und den im jeweiligen IRT-Modell spezifizierten Itemeigenschaften, die die Itemantworten beeinflussen. Diese Annahme ist verletzt, wenn neben der Personenfähigkeit ein Leistungsabfall die Itemantworten bedingt (vgl. De Boeck, Cho & Wilson, 2011; Yen, 1993).

Wie im vorherigen Kapitel 1 erläutert wurde, sind Not-Reached-Items und Leistungsabfall in LSAs eher die Regel als die Ausnahme. Für beide Fälle wurde daher in verschiedenen Studien die Auswirkungen auf die Parameterschätzung untersucht und Modelle entwickelt, um die Phänomene abzubilden und ihr Ausmaß zu bestimmen sowie den Zusammenhang mit der Personenfähigkeit zu schätzen.

2.3 Umgang mit Not-Reached-Items

Die Bedeutung, die das Auftreten von Not-Reached-Items auf die Schätzung von Personen- und Itemparametern hat, kann aus zwei Perspektiven betrachtet werden. Erstens ergibt sich die Frage nach der angemessenen Behandlung der fehlenden Itemantworten (vgl. Lord, 1974, 1980). Üblicherweise werden Not-Reached-Items entweder als falsche Antwort kodiert oder als fehlender Wert behandelt. Hier haben sich in den verschiedenen LSAs unterschied-

liche Vorgehensweisen etabliert. Fehlende Itemantworten für die Not-Reached-Items stellen eine Form von *missing data* dar (vgl. Glas & Pimentel, 2008; Pohl & Carstensen, 2013). Eine wichtige Frage für *missing data* ist die nach der *Ignorierbarkeit*, das heißt inwieweit die Daten zufällig fehlen oder dem Fehlen eher eine Systematik zugrunde liegt, die zu falschen Rückschlüssen über die Verteilung der Variable führen kann (vgl. Lüdtke, Robitzsch, Trautwein & Köller, 2007; Schafer & Graham, 2002). Zweitens ist daher von Interesse, ob Not-Reached-Items der Kategorie *ignorierbar fehlende Werte* zugeordnet werden können.

2.3.1 Kodierung von Not-Reached-Items

Es haben sich in den nationalen und internationalen LSAs verschiedene Umgangsweisen für Not-Reached-Items etabliert: PISA hat bis 2015 ein zweistufiges Verfahren verwendet, nach dem Not-Reached-Items für die Schätzung der Itemparameter als fehlender Wert, in der Schätzung der Personenfähigkeiten als falsch bewertet werden (z. B. OECD, 2012). Ein ähnliches Verfahren wird in TIMSS und PIRLS angewendet (Foy & Yin, 2017; Martin et al., 2016). Im aktuellen PISA-Zyklus von 2015 werden Not-Reached-Items in beiden Fällen als fehlende Werte behandelt, sodass für Personen- wie Itemparameter dasselbe Messmodell zugrunde gelegt wird (OECD, 2017). Auch in NAEP und NEPS werden Not-Reached-Items in beiden Fällen als fehlende Werte behandelt (NCES, 2016; Pohl & Carstensen, 2012).

Die verschiedenen Kodierweisen von Not-Reached-Items können zu unterschiedlichen Ergebnissen bezüglich der Verteilung der Personenfähigkeiten führen: Normalerweise wird die Kodierung als falsche Antwort zu geringeren Fähigkeitsscores führen[8] und, wenn Personenfähigkeit und Testabbruch zusammenhängen, wird sich die Verteilung der Scores dadurch systematisch verändern (Ludlow & O'Leary, 1999).

[8] Die Personenfähigkeit wird tendentiell unterschätzt, da durch die Behandlung als falsche Antwort für alle unbeobachteten Itemantworten nicht berücksichtigt wird, dass eine Person eine richtige Antwort hätte geben können (Pohl et al., 2014).

In verschiedenen Simulationsstudien wurden die Auswirkungen der Kodierung der Not-Reached-Items auf die Parameterschätzung untersucht. Die Ergebnisse zeigen, dass die Behandlung der Not-Reached-Items als falsche Antwort zu einem Bias in der Schätzung führt (Finch, 2008; Oshima, 1994; Pohl et al., 2014).[9] Daher wurde die Kodierung als falsche Antwort häufig kritisiert (z. B. Pohl & Carstensen, 2013; Rose, 2013; Rose, von Davier & Nagengast, 2017).[10] Stattdessen schlagen etwa Pohl und Carstensen (2013) vor, Not-Reached-Items als fehlende Werte zu behandeln. Bei der Behandlung als fehlender Wert kann es allerdings zu einer Überschätzung der Personenfähigkeit kommen, weil bei der Parameterschätzung auf Basis der unvollständigen Daten für jede fehlende Itemantwort implizit eine Wahrscheinlichkeit größer Null angenommen wird (Robitzsch, 2016; Rost, 2004).

Auch Robitzsch (2016) untersucht die Auswirkungen der Kodierung von Not-Reached-Items. Robitzsch (2016) betrachtet die Kodierung als falsche Antwort als den einen Extrempunkt und die Kodierung als fehlender Wert als den anderen Extrempunkt (vgl. Rost, 2004), die beide zu einer Verschätzung der Personenfähigkeit führen können. Robitzsch (2016) zeigt in einer Anwendung auf Daten aus PIRLS 2011, dass die verschiedenen Ansätze (Kodierung als falsche Antwort/ fehlender Wert und Abstufungen dazwischen) zu unterschiedlichen Schätzungen für die Rangreihe der Ländermittelwerte führt.

Darüber hinaus implizieren die Kodierung als falsche oder fehlende Antwort jeweils eine bestimmte Sichtweise auf die zu erfassende Personenfähigkeit, nämlich entweder als die Fähigkeit, Items in einer festgelegten Zeitspanne richtig zu lösen, oder als die Fähigkeit, Items unabhängig von der Bearbeitungszeit richtig zu lösen (Lord, 1980; Rohwer, 2013;

[9] Es ist zu beachten, dass in den zitierten Studien die Daten jeweils so generiert wurden, dass nicht alle fehlenden Itemantworten falsch sind. In so einem Fall würde die Behandlung als falsche Antwort zu unverzerrten Schätzern führen, weil das Auswertungs- mit dem datengenerierenden Modell identisch ist (vgl. Lord, 1980, 1983; Robitzsch, 2016; Rohwer, 2013; s. Fußnote 8).

[10] Zur ausführlichen Kritik an der Kritik siehe aber Robitzsch (2016, Kap. 6).

Rost, 2004). Der Umgang mit Not-Reached-Items richtet sich also auch nach der zugrunde gelegten Definition von Personenfähigkeit.[11]

Werden Not-Reached-Items als fehlende Werte kodiert, kann in das Messmodell eine *Missingness-Variable* integriert werden, die das Ausmaß von fehlenden Itemantworten beschreibt. Dies erlaubt, die gemeinsame Verteilung von fehlenden und beobachteten Itemantworten zu modellieren und auf einen Zusammenhang zwischen beiden zu untersuchen. Auf Basis der Missing-Data-Literatur (vgl. Rubin, 1976) wurden verschiedene modellbasierte Behandlungsweisen für Not-Reached-Items vorgeschlagen (z. B. Debeer et al., 2017; Glas & Pimentel, 2008; Holman & Glas, 2005; Rose, 2013), die in Abschnitt 2.3.3, nach einer Einbindung in den *Missing-Data*-Kontext im folgenden Abschnitt 2.3.2, besprochen werden.

2.3.2 Kategorisierung von fehlenden Werten

Aufbauend auf den Arbeiten von Rubin (1976) unterscheiden Schafer und Graham (2002) drei Kategorien von fehlenden Werten: *missing completely at random* (MCAR), *missing at random* (MAR) und *missing not at random* (MNAR; s. Schafer & Graham, 2002, S. 151; vgl. Little & Rubin, 2002; Rubin, 1976). Im Kontext von Item-Response-Modellen hängt die Kategorisierung nach MCAR, MAR und MNAR für fehlende Itemantworten für die Not-Reached-Items von dem Zusammenhang zwischen beobachteten Itemantworten, latenter Personenfähigkeit und dem Missingness-Prozess ab (vgl. Mislevy & Wu, 1996). Im Folgenden werden die Kategorien für fehlende Itemantworten für die Not-Reached-Items beschrieben und es wird angenommen, dass es keine anderen Formen fehlender Itemantworten gibt. Die Kategorisierung ist aber analog gültig für ausgelassene Items oder andere Formen fehlender Daten (vgl. Bradlow & Thomas, 1998; Holman & Glas, 2005).

[11] Für ausführliche Besprechungen zum Umgang mit Not-Reached-Items sei auf Rohwer (2013) und auf die Arbeiten von Rose et al. (Rose, 2013; Rose, von Davier & Nagengast, 2015; Rose et al., 2017) verwiesen.

Sei $\boldsymbol{X}_{komp}$ die Matrix der kompletten Daten, die sich aus beobachteten und fehlenden Itemantworten zusammensetzt, $\boldsymbol{X}_{komp} = (\boldsymbol{X}_{beob}, \boldsymbol{X}_{mis})$ (vgl. Lüdtke et al., 2007; Rose et al., 2010). Ferner sei $\boldsymbol{D}$ die Menge von *Missing-Indikatoren*, mit $d_{pi} \in \boldsymbol{D}$ für jede Person p und jedes Item i und $d_{pi} = 0$, wenn Person p Item i bearbeitet hat, $d_{pi} = 1$, wenn Item i das erste Not-Reached-Item von Person p ist, und $d_{pi} = \text{NA}$ für alle nachfolgenden Not-Reached-Items, wobei NA einen fehlenden Wert bezeichnet (vgl. Glas & Pimentel, 2008).[12] Die latente Personenvariable θ sei über $\boldsymbol{X}_{komp}$ definiert und sei $\boldsymbol{Z}$ eine Menge von beobachteten Kovariaten.

Fehlende Itemantworten sind MCAR, wenn ihr Fehlen in keinem Zusammenhang mit $\boldsymbol{X}_{beob}$, $\boldsymbol{X}_{mis}$ und $\boldsymbol{Z}$ steht, das bedeutet, dass diese Werte „vollständig zufällig“ fehlen (Lüdtke et al., 2007, S. 104), das heißt: $P(\boldsymbol{D}|\boldsymbol{X}_{beob}, \boldsymbol{X}_{mis}, \boldsymbol{Z}, \theta) = P(\boldsymbol{D})$ (vgl. Glas & Pimentel, 2008; Rose et al., 2015). Fehlende Itemantworten sind MAR, wenn ihr Fehlen nur von den beobachteten Daten abhängt: $P(\boldsymbol{D}|\boldsymbol{X}_{beob}, \boldsymbol{X}_{mis}, \boldsymbol{Z}, \theta) = P(\boldsymbol{D}|\boldsymbol{X}_{beob}, \boldsymbol{Z})$ (vgl. Glas & Pimentel, 2008; Kuha, Katsikatsou & Moustaki, 2018; Rose et al., 2015). Unter Kontrolle dieser beobachteten Daten ist das Fehlen somit unsystematisch. Trifft keine der beiden Bedingungen zu, sind die fehlenden Werte MNAR: Ihr Fehlen hängt damit von den fehlenden Werten $\boldsymbol{X}_{mis}$ oder der latenten Variable θ ab (vgl. Glas & Pimentel, 2008; Kuha et al., 2018; Rose et al., 2015).

Da fehlende Werte unter MCAR- oder MAR-Bedingungen nach Kontrolle der beobachteten Daten nicht systematisch fehlen, werden sie auch als *ignorierbar fehlende Werte* bezeichnet (vgl. Enders, 2010). Im Gegensatz dazu handelt es sich bei MNAR um *nicht-ignorierbar fehlende Werte* (Lüdtke et al., 2007; Schafer & Graham, 2002). Der *Missingness-Prozess*, also das Zustandekommen eines fehlendes Wertes, muss bei nicht-ignorierbar fehlenden Werten im Modell berücksichtigt werden, um Verzerrungen in den Parameterschät-

[12] Gleichermaßen kann $\boldsymbol{D}$ über die Anzahl der Not-Reached-Items R definiert werden, da Not-Reached-Items vollständig durch ihre Anzahl oder alternativ die Itemposition des ersten Auftretens bestimmt sind.

zungen zu vermeiden (Bradlow & Thomas, 1998; Little & Rubin, 2002; Mislevy & Wu, 1988, 1996; Schafer & Graham, 2002).

Die Klassifizierung nach MCAR, MAR und MNAR ist abhängig von den beobachteten Daten (Graham, 2009). Im Allgemeinen sind Not-Reached-Items MNAR, weil sie von der latenten Variable abhängen (vgl. Glas & Pimentel, 2008; Holman & Glas, 2005; Rose et al., 2010). Damit sind die fehlende Itemantworten für Not-Reached-Items nicht-ignorierbar und sollten im Modell berücksichtigt werden (Mislevy & Wu, 1996; Schafer & Graham, 2002).

2.3.3 Modelle für Not-Reached-Items

Anders als ausgelassene Items kommen Not-Reached-Items definitionsgemäß immer als Block am Testende vor, sodass nach dem Auftreten des ersten Not-Reached-Items die weiteren determiniert sind (vgl. Rose et al., 2015). Ein ähnliches Muster von fehlenden Werten ergibt sich in Survivalanalysen, die den Zeitpunkt des Ausscheidens zum Untersuchungsgegenstand haben (vgl. Allison, 2014). Im Folgenden werden zunächst Ansätze zur Modellierung der gemeinsamen Verteilung von beobachteten und fehlenden Itemantworten erläutert (Abschnitt 2.3.3.1). Diese Ansätze sind nicht spezifisch für Not-Reached-Items, sondern können allgemein auf fehlende Werte und zum Beispiel auch im Kontext von ausgelassenen Itemantworten angewandt werden. Im nachfolgenden Abschnitt 2.3.3.2 wird das Konzept der Survivalanalyse auf Not-Reached-Items übertragen.

2.3.3.1 Gemeinsame Verteilung beobachteter und fehlender Itemantworten

Es existieren verschiedene Ansätze, um die gemeinsame Verteilung von beobachteten Werten und nicht-ignorierbar fehlenden Itemantworten zu modellieren. Seien $\boldsymbol{X}_{komp}$ und $\boldsymbol{D}$, wie oben definiert, die Mengen der kompletten Itemantworten und der Missing-Indikatoren. Die vorgeschlagenen Ansätze modellieren die gemeinsame Verteilung von $\boldsymbol{X}_{komp}$ und $\boldsymbol{D}$,

$P(\boldsymbol{X}_{komp}, \boldsymbol{D})$, wobei entweder $\boldsymbol{X}_{komp}$ als abhängige Variable gegeben $\boldsymbol{D}$: $P(\boldsymbol{X}_{komp}, \boldsymbol{D}) = P(\boldsymbol{X}_{komp}|\boldsymbol{D})P(\boldsymbol{D})$ (*Pattern-Mixture-Modelle*: Little, 1995; vgl. Harel & Schafer, 2009; Kuha et al., 2018; O'Muircheartaigh & Moustaki, 1999) oder $\boldsymbol{D}$ als abhängige Variable gegeben $\boldsymbol{X}_{komp}$: $P(\boldsymbol{X}_{komp}, \boldsymbol{D}) = P(\boldsymbol{D}|\boldsymbol{X}_{komp})P(\boldsymbol{X}_{komp})$ (*Selection-Modelle*: Diggle & Kenward, 1994; vgl. Harel & Schafer, 2009; Kuha et al., 2018; O'Muircheartaigh & Moustaki, 1999) betrachtet werden (z. B. B. Muthén, Asparouhov, Hunter & Leuchter, 2011; Rose et al., 2017).[13] Pattern-Mixture-Modelle und Selection-Modelle unterscheiden sich darin, ob $\boldsymbol{D}$ als Prädiktor oder als Kriterium modelliert wird, und haben somit unterschiedliche Einsatzbereiche.

Formen von Pattern-Mixture-Modellen können über einen regressionbasierten Ansatz realisiert werden, in dem die Anzahl der Not-Reached-Items R (oder eine Funktion dieser) als Prädiktor für θ berücksichtigt wird (z. B. Köhler et al., 2015a; Pohl et al., 2014; Rose et al., 2015), oder über ein Mehrgruppenmodell, in dem jede Gruppe durch ein Missingness-Muster spezifiziert ist (z. B. Rose et al., 2010). Die Schätzung einer Regression von θ auf R kann um Polynome höherer Ordnung erweitert werden, um komplexe, nichtlineare Zusammenhänge abzubilden (Rose et al., 2015). Ebenso können im Mehrgruppenmodell je Ausprägung von R unterschiedliche Regressionskoeffizienten geschätzt werden (Rose et al., 2010). Kovariaten $\boldsymbol{Z}$ können in Pattern-Mixture-Modellen als weitere Prädiktoren berücksichtigt werden (z. B. Köhler et al., 2015a; Rose et al., 2017).

Anders als in den Pattern-Mixture-Modellen wird in Selection-Modellen die Verteilung von $\boldsymbol{D}$ konditional zu $\boldsymbol{X}_{komp}$ (und $\boldsymbol{Z}$) betrachtet. Selection-Modelle wurden bisher kaum

[13] Üblicherweise wird die Verteilung von $\boldsymbol{X}_{komp}$ und $\boldsymbol{D}$ konditional zu Kovariaten $\boldsymbol{Z}$, der latenten Fähigkeitsvariable θ und weiteren Parametern für das Messmodell von θ und für den Missingness-Prozess betrachtet (vgl. Rose et al., 2015). Auf den Einbezug von $\boldsymbol{Z}$, θ und ggf. weiteren Parametern für das Messmodell von θ und für den Missingness-Prozess (vgl. Rose et al., 2015) wird an dieser Stelle verzichtet, um die Unterschiede in beiden Modellierungsansätzen deutlicher hervorzuheben. Ausführlichere Darstellungen von Pattern-Mixture- und Selection-Modellen im Kontext von IRT und Strukturgleichungsmodellen finden sich zum Beispiel bei Harel und Schafer (2009), Kuha et al. (2018), O'Muircheartaigh und Moustaki (1999) und Rose et al. (2015).

im IRT-Kontext angewendet (z. B. Rose, 2013). Der Vorteil von Selection-Modellen ist, dass der Missingness-Prozess als abhängige Variable betrachtet wird. Während Pattern-Mixture-Modelle vor allem angewendet werden, um einen Bias bei der Schätzung der Personenfähigkeit zu reduzieren, der durch MNAR-Bedingungen entsteht (z. B. Pohl et al., 2014), ermöglichen Selection-Modelle, gezielt Prädiktoren des Missingness-Prozesses zu untersuchen. Übertragen auf Not-Reached-Items kann man mit Selection-Modellen der Frage nachgehen, *wer* den Test abbricht, das heißt, wie Testabbruch durch Personenfähigkeit und Kovariaten erklärt werden kann.

Ein dritter Modelltyp sind *Shared-Parameter-Modelle* (vgl. Albert & Follmann, 2009; Harel & Schafer, 2009; Wu & Carroll, 1988), in denen die Kovarianz zwischen der latenten Fähigkeitsvariable θ (definiert über $\boldsymbol{X}_{komp}$) und einer latenten Missingness-Variable δ (definiert über $\boldsymbol{D}$ bzw. über R) betrachtet wird. Dabei wird normalerweise angenommen, dass θ und δ bivariat normalverteilt sind.

Glas und Pimentel (2008) haben ein Modell für Not-Reached-Items vorgeschlagen, in dem über die Missing-Indikatoren eine latente Missingness-Variable δ für den Testabbruch definiert ist. Besteht eine signifikante Korrelation zwischen θ und δ, deutet dies darauf, dass fehlende Werte auf den Not-Reached-Items MNAR sind. Ein ähnliches Shared-Parameter-Modell wurde von Holman und Glas (2005) für ausgelassene Items entwickelt. Dieses Modell wurde von Glas, Pimentel und Lamers (2015) um Kovariaten für θ und δ erweitert.[14] Köhler, Pohl und Carstensen (2015b) formulieren dieses Modell zu einem *General Diagnostic Model* (GDM; von Davier, 2008) um, um nichtlineare Beziehungen zwischen θ und δ abbilden zu können. Mithilfe eines GDM können beliebige Verteilungen latenter Variablen dargestellt werden, indem nur diskrete Stützstellen der Verteilungen geschätzt werden. Damit lösen Köhler et al. (2015b) die Annahme der bivariaten Normalverteilung von θ und

[14] Diese Missingness-Variable δ wird über Missing-Indikatoren für die ausgelassenen Items definiert.

δ auf. Allerdings schätzen Köhler et al. (2015b) nur den Zusammenhang von θ und δ und berücksichtigen keine Kovariaten.

2.3.3.2 Survivalanalyse des Testabbruchs

Das erste Item im Test, das eine Person nicht mehr bearbeitet, das heißt der Beginn der Not-Reached-Items, kann als ein nicht wiederkehrendes Ereignis betrachtet werden. Die Ereignisdaten- oder Survivalanalyse für diskrete Messzeitpunkte modelliert den Zusammenhang zwischen dem Zeitpunkt, an dem ein nicht wiederkehrendes Ereignis eintritt, und stabilen wie auch zeitlich veränderlichen Personenmerkmalen (Allison, 2014; Singer & Willett, 1993). Genauer geht es in der Survivalanalyse um Korrelate der Wahrscheinlichkeit für das Eintreten des Ereignisses zu einem bestimmten Zeitpunkt. Die Wahrscheinlichkeit, dass für eine bestimmte Person das Ereignis zu einem bestimmten Zeitpunkt eintritt, gegeben, dass es bisher nicht eingetreten ist, wird als *Hazard* bezeichnet. Der Zusammenhang zwischen Hazard und Personenmerkmal kann über eine logistische Regression von Personenmerkmal auf den Logit des Hazards geschätzt werden (für einen Überblick s. Allison, 2014).

Das Modell von Glas und Pimentel (2008) weist Ähnlichkeiten mit einer Survivalanalyse für diskrete Messzeitpunkte auf, da hier der Zusammenhang von Testabbruch an einer bestimmten (diskreten) Itemposition und Personenfähigkeit modelliert wird. Über den Beginn der Not-Reached-Items, das heißt den Testabbruch, wird eine latente Missingness-Variable definiert und deren Kovarianz mit der Personenfähigkeit geschätzt. Glas und Pimentel (2008) greifen für ihr Modell auf das *Steps-Modell* von Verhelst, Glas und de Vries (1997) zurück: Im Steps-Modell wird die Wahrscheinlichkeit geschätzt, eine bestimmte Stufe eines mehrstufigen Items zu erreichen, gegeben, dass alle bisherigen Stufen erreicht wurden. Das Konzept der Stufen übertragen Glas und Pimentel (2008) auf die Items eines Tests, um

die Wahrscheinlichkeit zu schätzen, dass ein Item erreicht (d. h. bearbeitet) wird, gegeben, dass alle vorherigen Items erreicht wurden.

Anders als in einer Survivalanalyse schätzen Glas und Pimentel (2008) nicht den direkten Effekt von Personenfähigkeit auf den Testabbruch, sondern definieren eine latente Missingness-Variable, die über die NRIs repräsentiert ist. Der Zusammenhang von Testabbruch und Personenfähigkeit zeigt sich in der Korrelation beider latenter Variablen. Das Messmodell für die Missingness-Variable entspricht einem 1PL mit Restriktionen auf den Schwierigkeitsparametern der Missing-Indikatoren, sodass die Schwierigkeitsparameter als lineare Funktion über die Itemposition geschätzt werden. Glas und Pimentel (2008) nehmen an, dass die Wahrscheinlichkeit, ein Item nicht mehr zu erreichen, im Testverlauf monoton ansteigt. Ferner wird der Zusammenhang zwischen Personenfähigkeit und Testabbruch als linear und die gemeinsame Verteilung beider latenten Variablen als bivariat normalverteilt angenommen. Diese beiden Annahmen, der monotone Anstieg in den Hazards und der lineare Zusammenhang von Testabbruch und Personenfähigkeit sind Restriktionen, die für Survivalanalysen diskreter Messzeitpunkte nicht erforderlich sind.

Das Modell von Glas und Pimentel (2008) impliziert, dass der Testabbruch im Testverlauf immer wahrscheinlicher wird – jedoch kann es Situationen geben, wo Testteilnehmende entweder bereits zu Beginn des Tests oder aber erst am Testende die Bearbeitung abbrechen: Zum Beispiel können Personen mit einer sehr geringen Testbearbeitungsmotivation schon früh aufhören, den Test zu bearbeiten, während motivierte Testteilnehmende am Testende unter Zeitdruck die letzten Items nicht mehr erreichen. Sind die Gruppen der motivierten und wenig motivierten Testteilnehmenden gleich groß, wird sich das in einem u-förmigen Hazardverlauf zeigen.

Typisch für Survivalanalysen diskreter Messzeitpunkte ist, dass für die Prädiktoren ein über die Zeit variierender Effekt auf den Abbruch modelliert werden kann (Allison, 2014; B. Muthén & Masyn, 2005). Übertragen auf Not-Reached-Items bedeutet dies, dass der Zu-

sammenhang mit der Personenfähigkeit am Testende anders als am Testbeginn sein kann. Beispielsweise kann der Einfluss der Personenfähigkeit auf den Testabbruch im mittleren Testbereich am stärksten ausfallen, wenn mit verschiedenen Ausprägungen von Personenfähigkeit verschieden zeitintensive Testbearbeitungsstrategien einhergehen: Ein sehr früher Testabbruch ist für alle Ausprägungen der Personenfähigkeit unwahrscheinlich, sodass am Testanfang kein Zusammenhang zwischen Testabbruch und Fähigkeit besteht. Personen, die sehr langsam arbeiten, erreichen dann aber nur den mittleren Bereich des Tests, während Personen mit schnelleren Strategien das Testende erreichen. Denkbar ist, dass ein langsames Arbeiten mehr Gründlichkeit bedeutet und zu höheren Fähigkeitsscores führt oder dass Personen mit höherer Fähigkeit schneller die richtigen Itemlösungen wissen. Die im Modell von Glas und Pimentel (2008) definierte Korrelation zwischen Personenfähigkeit und Missingness-Variable impliziert einen konstanten Effekt von Personenfähigkeit auf den Testabbruch und ermöglicht nicht, einen positionsvariierenden Effekt zu schätzen.

Darüber hinaus sind auch nichtlineare Effekte denkbar: Sowohl niedrige wie auch hohe Werte auf der Fähigkeitsvariable können mit einer höheren Wahrscheinlichkeit für den Testabbruch assoziiert sein, etwa wenn Personen mit geringer Fähigkeit durch das zu schwere Testmaterial frustriert den Test mit höherer Wahrscheinlichkeit abbrechen, und gleichzeitig Personen durch sehr gründliches Arbeiten einen hohen Fähigkeitsscore erhalten, aber ebenfalls mit höherer Wahrscheinlichkeit aus Zeitmangel das Testende nicht erreichen. In diesem Fall wäre für Personen mit einem mittleren Fähigkeitsniveau die Wahrscheinlichkeit, das Testende zu erreichen, am höchsten. Ein Modell für lineare Beziehungen kann einen solchen Effekt nicht aufdecken und könnte im Extremfall die Unabhängigkeit beider Variablen implizieren, wenn sich die Effekte für hohe und niedrige Werte der Personenfähigkeit gegenseitig aufheben.

2.4 Leistungsabfall als Störvariable

Der Leistungsabfall kann als eine weitere Personenvariable betrachtet werden, die neben der Personenfähigkeit das Antwortverhalten auf den Items am Testende bedingt. Wie oben erläutert, verletzt dies eine der Grundannahmen der IRT, weil die Itemantworten nicht allein auf die Personenfähigkeit zurückzuführen sind. Dies zeigt sich, wenn der Leistungsabfall nicht berücksichtigt wird, in lokalen Abhängigkeiten der betroffenen Items am Testende (Lu & Sireci, 2007; Yen, 1993). Leistungsabfall kann somit als eine nichtindentierte Dimension, eine Bias- oder Störvariable betrachtet werden, die der Test miterfasst (De Boeck et al., 2011; Kok, 1988; Hambleton & Swaminathan, 1985; van der Linden, 2007b; Oshima, 1994), was zu konstruktirrelevanter Varianz führt (Eklöf, 2007; Haladyna & Downing, 2004; Lu & Sireci, 2007). Wenn der Leistungsabfall nicht berücksichtigt wird, schränkt dies die Validität des Tests ein (Eklöf, 2010; Haladyna & Downing, 2004; van der Linden, 2007b; Lord, 1980; Lu & Sireci, 2007; Weirich, 2015; S. L. Wise & DeMars, 2005), und kann zu einem Bias in den Parameterschätzungen führen (De Boeck et al., 2011; Oshima, 1994; Weirich, 2015; S. L. Wise, Kingsbury, Thomason & Kong, 2004; Suh, Cho & Wollack, 2012).

2.4.1 Auswirkungen auf die Parameterschätzung

Da sich der Leistungsabfall in der verringerten Wahrscheinlichkeit einer richtigen Itemantwort zeigt, können die Item- oder Personenparameter des IRT-Modells verschätzt werden. Items an späteren Positionen erscheinen durch die verringerte Lösungswahrscheinlichkeit schwieriger (van Barneveld, 2003; Bolt et al., 2002; Davey & Lee, 2011) und die Itemschwierigkeiten können deswegen überschätzt sein (z. B. Le, 2007; Yen, 1980). Bei den Itemdiskriminationen sind sowohl Über- wie auch Unterschätzungen aufgrund eines nichtberücksichtigten Leistungsabfalls denkbar (Frey, Bernhardt & Born, 2017), empirisch finden sich relativ zum Wert an früheren Itempositionen sowohl eine Überschätzung (Yen, 1980) als

auch eine Unterschätzung für Items am Testende (G. Nagy, Nagengast, Frey, Becker & Rose, 2016). Hingegen finden De Boeck et al. (2011) und Le (2007) zwar auch Unterschiede in den Diskriminationsparametern, aber die Differenzen der Parameter zu Beginn und am Testende weisen keine eindeutige Über- oder Unterschätzung aus. Zusammengefasst deuten jedoch auch die Befunde zur Itemdiskrimination allgemein auf eine Verschätzung hin, wenn der Leistungsabfall nicht berücksichtigt wird (vgl. Oshima, 1994). Die Verzerrung der Itemparameter kann sich auf die Schätzung der Personenfähigkeit auswirken und hier zu einer Überschätzung für Personen ohne Leistungsabfall führen, die vermeintlich schwierige Items lösen (Yang, 2007) oder zu einer Unterschätzung für Personen mit Leistungsabfall, die die Items am Testende mit geringerer Wahrscheinlichkeit lösen (Davey & Lee, 2011).

Die Verschätzung der Itemparameter stellt ein Problem fürs *Linking* (vgl. Dorans, Pommerich & Holland, 2007; Kolen & Brennan, 2004; Wu, 2010) dar. Linking bedeutet im allgemeinen, dass eine funktionale, eindeutige Beziehung zwischen Testwerten (*Scores*) aus verschiedenen Tests hergestellt wird, sodass Scores aus dem einen Test mit denen aus dem anderen Test vergleichbar sind (Mislevy, Johnson & Muraki, 1992).[15] Mit einem Linking auf Basis der IRT werden die Itemparameter aus verschiedenen Testerhebungen auf eine gemeinsame Metrik gebracht, um darüber eine gemeinsame Skala für die Fähigkeitsscores festzulegen (Kolen, 2004; Kolen & Brennan, 2004). Die zu verlinkenden Tests können dabei aus vollständig identischen Itemmengen bestehen oder eine identische Teilmenge von Items (Ankeritems) beinhalten. Die Validität des Linking hängt davon ab, inwieweit die Itemparameter in den verschiedenen Erhebungen unverzerrt geschätzt werden.

Die durch den Leistungsabfall bedingte Verzerrung der Itemparameter kann die Validität des Linking bedrohen (Bolt et al., 2002; Mittelhaëuser, Béguin & Sijtsma, 2013; Wollack, Cohen & Wells, 2003). Dies gilt besonders dann, wenn sich die Linkingstichproben im Ausmaß ihres Leistungsabfalls unterscheiden (Mittelhaëuser et al., 2015a; Mittel-

[15] Es gibt viele unterschiedliche Ansätze zum Linking, innerhalb und außerhalb der IRT. Eine ausführliche Darstellung verschiedener Verfahren findet sich bei Kolen und Brennan (2004).

haëuser, Béguin & Sijtsma, 2015b; Wolf, Smith & Birnbaum, 1995; für einen Überlick s. Thissen, 2007) oder wenn die Ankeritems an verschiedenen Positionen in den Tests vorkommen und dadurch in unterschiedlichem Maße vom Leistungsabfall betroffen sind (Bolt et al., 2002; Wollack et al., 2003). Ferner können Gruppenunterschiede in der Testleistung oder die Leistungsentwicklung verzerrt geschätzt werden, wenn der Leistungsabfall nicht berücksichtigt wird (Goldhammer, Martens, Christoph & Lüdtke, 2016; Mittelhaëuser et al., 2015b; G. Nagy, Retelsdorf et al., 2017; S. L. Wise & DeMars, 2010).

Auch bei der Skalierung, der Berechnung von Scores (vgl. Kolen, Tong & Brennan, 2011), auf Basis von Itemparametern, die in einer anderen Erhebung berechnet wurden, ist ein unberücksichtigter Leistungsabfall ein Problem (van Barneveld, 2007; Mittelhaëuser et al., 2015a; Wolf et al., 1995). Die Itemkalibrierung (d. h. die Schätzung der Itemparameter) für High-Stakes-Tests wird meist in Low-Stakes-Prätests durchgeführt (Davey & Lee, 2011; Mittelhaëuser, Béguin & Sijtsma, 2011; Mittelhaëuser et al., 2013). Da Low-Stakes-Tests keine Konsequenzen für die Testteilnehmenden haben, ist anzunehmen, dass das Ausmaß des Leistungsabfalls in der Low-Stakes-Testsituation aufgrund geringer Testbearbeitungsmotivation höher ausfällt als im High-Stakes-Test, was wiederum dazu führen kann, dass die Parameter der Items am Testende systematisch unterschiedlich geschätzt werden. Werden die Itemparameter aus dem Low-Stakes-Test zur Schätzung der Personenfähigkeiten im High-Stakes-Test verwendet, können die Fähigkeitsscores verzerrt werden (Davey & Lee, 2011; van der Linden, 2011; Mittelhaëuser et al., 2011, 2015a, 2015b).

2.4.2 Modellierung des Leistungsabfalls

2.4.2.1 Modelle für Positionseffekte

In vielen Anwendungen wird der Leistungsabfall über den Vergleich der Itemparameter an frühen und späteren Positionen im Test bestimmt (z. B. Debeer & Janssen, 2013; Debeer

et al., 2014; Hartig & Buchholz, 2012; G. Nagy, Lüdtke & Köller, 2016). Um neben der Varianz des Leistungsabfalls auch den Mittelwert schätzen zu können, setzt dies aber ein Multi-Matrix-Design voraus, in dem Testhefte, zwischen denen die Positionen der Items variieren, zufällig auf die Stichprobe verteilt werden (s. Gonzalez & Rutkowski, 2010).

Um das Ausmaß des Leistungsabfalls anhand der Itempositionseffekte zu schätzen, wird in das IRT-Modell ein Parameter für die Änderung der Itemschwierigkeit zu späteren Positionen eingefügt. Die Annahme eines personenspezifischen Parameters ermöglicht, das indiviudelle Ausmaß von Leistungsabfall zu schätzen (Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Robitzsch, 2009) und mit anderen Personenmerkmalen in Beziehung zu setzen (Bulut et al., 2017).

Ein zweidimensionales IRT-Modell mit personenspezifischem Positionseffekt haben Debeer und Janssen (2013) und auch Hartig und Buchholz (2012) vorgeschlagen. In beiden Modellen wird neben der Personenfähigkeit eine latente Variable für den Positionseffekt modelliert, die die Veränderung der Lösungswahrscheinlichkeit an späteren Positionen im Test bedingt. In beiden Modellen wird für die Leistungsabfall-Variable eine Normalverteilung angenommen, womit implizit auch angenommen wird, dass es Testteilnehmende mit einem positiven Positionseffekt geben kann, für die die Items an späteren Positionen leichter statt schwieriger werden. Diese Annahme ist allerdings nicht erforderlich und es können (theoretisch) beliebige andere Verteilungen zugrunde gelegt werden.

In Testdesigns, in denen nicht alle Items an der ersten Position vorkommen oder die Itempositionen konstant gehalten werden, können diese Modelle nicht oder nur mit zusätzlichen Restriktionen eingesetzt werden. Eine Alternative stellen hier *Mischverteilungsmodelle* dar, die explizit die Veränderung des Testbearbeitungsverhaltens modellieren. Mischverteilungsmodelle wurden verschiedentlich zur Analyse von Speededness oder geringer Testbearbeitungsmotivation vorgeschlagen (Bolt et al., 2002; De Boeck et al., 2011; Goegebeur

et al., 2008; Yamamoto, 1995). Dieser Ansatz zur Modellierung von Leistungsabfall wird im Folgenden näher beschrieben.

2.4.2.2 Mischverteilungsmodelle für den Leistungsabfall

Im alternativen Ansatz wird das Ausmaß des Leistungsabfalls über die Veränderung im Testbearbeitungsverhalten geschätzt. Dabei wird angenommen, dass (1) ein Teil der Personen der Stichprobe keinen Leistungsabfall zeigt, sondern bis zum Testende mit maximaler Leistung arbeitet, und dass (2) für die Itemantworten nach Beginn des Leistungsabfalls ein anderes Messmodell gilt. Diese Grundideen können in einem Mischverteilungsmodell repräsentiert werden. Mischverteilungsmodelle zum Leistungsabfall dienen dazu, die bezüglich des Leistungsabfalls heterogene Stichprobe zu *entmischen*, um für die Substichprobe der Personen mit anhaltend maximaler Leistung unverzerrte Item- und Personenparameter zu erhalten.

Allgemeines Konzept von Mischverteilungsmodellen. Mischverteilungsmodelle stellen eine Kombination aus IRT-Modell und Latent-Class-Modell (LCM; vgl. Formann, 1984; Langeheine & Rost, 1988) dar, sie verknüpfen die Klassifikation von Personen in unterschiedliche Subgruppen (LCM), das heißt bezüglich einer qualitativen latenten Variable, mit der Verortung von Personen gemäß ihrer Ausprägung auf einer quantitativen latenten Variable (IRT-Modell) (Collins & Lanza, 2010; Langeheine & Rost, 1988; Rost & Langeheine, 1997).

Latent-Class-Modelle dienen der Klassifikation von Personen in homogene Teilgruppen, so genannte latente Klassen (vgl. Collins & Lanza, 2010; Formann, 1984; Langeheine & Rost, 1988). Für jede Person in der Stichprobe werden im Analyseprozess die Wahrscheinlichkeiten zur Zugehörigkeit jeder der latenten Klassen berechnet. Während die Klassengrößen im Zuge der Parameterschätzung geschätzt werden, wird die Anzahl der Klassen

normalerweise vorgegeben (Rost, 2004).[16] Die Anzahl der Klassen kann theoretisch begründet sein oder die optimale Klassenanzahl kann per Modellselektion anhand von Modellpassungskriterien bestimmt werden.[17]

Mischverteilungsmodelle wurden vorgeschlagen, um heterogenes Testbearbeitungsverhalten abzubilden (Mislevy & Verhelst, 1990), und wurden verschiedentlich auf die Modellierung von Leistungsabfall angewendet (Bolt et al., 2002; Jin & Wang, 2014; Yamamoto, 1995). Die oben genannten Grundideen bei der Modellierung von Leistungsabfall werden dadurch realisiert, dass eine latente Klasse für alle Personen mit maximaler Leistung definiert wird (Annahme 1, s. o.) und dann eine oder weitere latente Klassen spezifiziert werden, in denen sich das Testbearbeitungsverhalten ab einer bestimmten Itemposition verändert (Annahme 2, s. o.).

Die Veränderung des Testbearbeitungsverhaltens meint dabei einen Leistungsabfall, der zu geringeren Lösungswahrscheinlichkeiten der betroffenen Items führt. Die verschiedenen Mischverteilungsmodelle für den Leistungsabfall unterscheiden sich darin, wie der Leistungsabfall operationalisiert wird, was weiter unten näher beschrieben wird.

Daneben unterscheiden sich die existierenden Modelle in der Anzahl der latenten Klassen, die definiert werden. Es wird davon ausgegangen, dass der Leistungsabfall abrupt *nach* einer bestimmten Itemposition einsetzt und dass die Testteilnehmenden bis zu dieser Position mit maximaler Leistung den Test bearbeitet haben.[18] Dies bedeutet, dass alle Testteilnehmenden das erste Item mit maximaler Leistung bearbeiten und dass ein Leistungsabfall frühestens ab dem zweiten Item einsetzen kann.

[16] Im Kontext Bayesianischer Verfahren gibt es allerdings auch Ansätze, die erlauben, die Klassenzahl als Parameter zu schätzen (*Dirichlet process mixture models*; z. B. Ghahramani, Griffiths & Sollich, 2006; Vidotto, 2018). Eine Besprechung dieser Modelle würde in der vorliegenden Dissertation zu weit führen und so sei an dieser Stelle nur auf weitere Literatur verwiesen.

[17] Dies kann beispielsweise über Indices zum komparativen Modellfit oder die optimale Klassenanzahl kann über bestimmte Zusatzanforderungen (etwa eine Mindestgröße für die einzelnen Klassen) erfolgen. Als Fitindex wird das *Bayesian Information Criterion* (BIC, Schwarz, 1978) vorgeschlagen (Li, Cohen, Kim & Cho, 2009; Nylund, Asparouhov & Muthén, 2007).

[18] Es sei noch einmal erwähnt, dass eine wesentliche Annahme für alle Modelle dieser Arbeit ist, dass die Testteilnehmenden den Test in der vorgegebenen Reihenfolge bearbeiten.

Damit existiert für jede Person der Stichprobe ein *Schaltpunkt* (*switching point*, vgl. von Davier & Yamamoto, 2007), an dem sich das Testbearbeitungsverhalten ändert, das heißt, an welchem der Leistungsabfall einsetzt. Der Schaltpunkt entspricht der letzten Itemposition vor Beginn des Leistungsabfalls. Für Personen, die bis zum Testende mit maximaler Leistung arbeiten, wird der Schaltpunkt auf die letzte Itemposition fixiert. Die Einteilung der Testteilnehmenden in latente Klassen erfolgt anhand der Schaltpunkte, sodass jeder Schaltpunkt einer latenten Klasse entspricht. Der Vektor von Itemantworten jeder Person wird damit in zwei Teile zerlegt. Die Itemantworten im ersten Teil lassen sich durch ein Standard-IRT-Modell (z. B. Gl. 2.2 oder Gl. 2.3) beschreiben, während das Messmodell für die Itemantworten des zweiten Teils einen Leistungsabfall berücksichtigt. Im Folgenden wird dies in einer generellen Formulierung für ein Mischverteilungsmodell für den Leistungsabfall näher erläutert.

Generelles Mischverteilungsmodell für den Leistungsabfall. Jeder Person p wird ein Wert auf zwei latenten Variablen zugeordnet. Neben der Personenfähigkeit θ_p existiert die diskrete *Schaltpunkt-Variable* δ_p, die für jede Person die Itemposition markiert, nach der Leistungsabfall einsetzt. Der Schaltpunkt kann Werte von 1 bis I annehmen, wobei I für die letzte Itemposition steht, das heißt $\delta_p = 1, \dots I$. Sei i der Itemindex. Ist $\delta_p = i$, bedeutet dies, dass der Leistungsabfall nach Itemposition i einsetzt. Der Leistungsabfall setzt direkt nach der ersten Position ein für $\delta_p = 1$, und ist $\delta_p = I$, bedeutet dies, dass die Person bis zum Testende mit maximaler Leistung arbeitet.

Für das generelle Mischverteilungsmodell wird angenommen, dass sich die Itemantworten vor und auch nach dem Schaltpunkt mit einem 2PL (s. Gl. 2.3) beschreiben lassen, dessen Modellparameter jedoch variieren. Statt eines 2PL als grundlegendes Messmodell können auch andere IRT-Modelle verwendet werden.

Sei η_{pi} der Logit der Wahrscheinlichkeit einer richtigen Itemantwort. Das generelle Mischverteilungsmodell ist:

$$\eta_{pi} = \begin{cases} a_i \cdot (\theta_p - b_i), \text{ if } \delta_p \geq i, \\ \tilde{a}_i \cdot (\tilde{\theta}_p - \tilde{b}_i), \text{ if } \delta_p < i. \end{cases} \tag{2.4}$$

Vor dem Schaltpunkt ergibt sich die Wahrscheinlichkeit einer richtigen Itemlösung aus der Personenfähigkeit θ_p und den Itemparametern a_i für die Itemdiskrimination und b_i für die Itemschwierigkeit. Nach dem Schaltpunkt hängt die Wahrscheinlichkeit einer richtigen Itemlösung von der Personenvariable $\tilde{\theta}_p$ und den Itemparametern $\tilde{a}_i$ und $\tilde{b}_i$ ab. Die Interpretation dieser Modellparameter, speziell, ob die Personenvariable nach dem Schaltpunkt weiterhin als *Personenfähigkeit* interpretiert werden kann, hängt von weiteren Modellspezifikationen ab und variiert zwischen den in der Literatur vorgeschlagenen Modellen.

Die gemeinsame Verteilung $P(\theta, \delta)$ hängt von den Annahmen zum Zusammenhang beider Variablen ab. Im typischen IRT-Modell ist θ eine kontinuierliche Variable, δ ist definitionsgemäß diskret. Es kann nun zum Beispiel eine stetige Verteilung für eine Transformation $T(\delta)$ von δ definiert werden, um eine bivariate Normalverteilung $P(\theta, T(\delta))$ (vgl. Suh et al., 2012) zu schätzen.

Beide Variablen, θ und δ, können aber auch als voneinander unabhängig spezifiziert werden (vgl. Jin & Wang, 2014). Diese Annahme ist etwa dann plausibel, wenn Speededness unabhängig von der Personenfähigkeit auftritt (z. B. Hailey et al., 2012). Das bedeutet, dass die klassenspezifischen Verteilungen von θ nicht die Abstufungen von δ variieren, sondern beispielsweise stets derselben Normalverteilung folgen.

Die gebräuchlichste Form der Verteilungsspezifikation ist, die konditionale Verteilung $P(\theta|\delta)$ und die Klassenwahrscheinlichkeiten zu betrachten, das heißt $P(\theta, \delta) = P(\theta|\delta) \cdot P(\delta)$. Für $P(\theta|\delta)$ wird eine Normalverteilung mit klassenspezifischem Mittelwert und klassenspezifischer Varianz angenommen. Gibt es viele Ausprägungen von δ, bietet es sich an,

Mittelwerte und Varianzen zu restringieren, um ein sparsameres Modell zu erhalten. Zum Beispiel schlagen Yamamoto und Everson (1997) vor, die Mittelwerte als lineare Funktion in Abhängigkeit von δ zu schätzen und die Varianzen konstant zu halten. Wenn es viele latente Klassen gibt, ist es auch für die Verteilung der Klassenwahrscheinlichkeiten $P(\delta)$ sinnvoll, diese anhand einer Funktion zu restringieren. Cao und Stokes (2008) haben vorgeschlagen, die Klassenwahrscheinlichkeiten der latenten Klassen für $\delta < I$, also die Klassen mit Leistungsabfall, als monoton steigende Funktion in Abhängigkeit der Größe der latenten Klasse für $\delta = I$ (d. h. die Klasse ohne Leistungsabfall) und eines Formparameters zu schätzen. Die vorgeschlagene Funktion lässt trotz der geringen Parameterzahl viele unterschiedliche Formen monoton steigender Kurven zu (vgl.Jin & Wang, 2014). Zugrunde liegt dabei die Annahme, dass der Beginn von Leistungsabfall zum Testende hin immer wahrscheinlicher wird. Für eine genaue Darstellung der Funktion sei auf Cao und Stokes (2008) und Jin und Wang (2014) verwiesen.

Spezifische Mischverteilungsmodelle für den Leistungsabfall. Es wurden verschiedene Mischverteilungsmodelle für den Leistungsabfall entwickelt. Das Modell von Yamamoto (1995) und das Modell von Bolt et al. (2002) sind dabei die bekanntesten und am häufigsten verwendeten (z. B. Boughton & Yamamoto, 2007; Cao & Stokes, 2008; De Boeck et al., 2011; Hailey et al., 2012; Mittelhaëuser et al., 2013, 2015a; Suh et al., 2012; Wollack et al., 2003; Yamamoto & Everson, 1995, 1997). Ein neueres Modell stammt von Jin und Wang (2014), das mit den beiden anderen Modellen gewisse Ähnlichkeiten aufweist. Gemeinsam ist allen drei Modellen, dass sie als Spezialfälle des oben vorgestellten generellen Mischverteilungsmodell für den Leistungsabfall (Gl. 2.4) formuliert werden. Die drei Mischverteilungsmodelle unterscheiden sich (1) in der Anzahl der latenten Klassen, (2) in der Interpretation der Personenvariable $\tilde{\theta}$ und (3) in der Spezifikation der Itemparameter (vgl. Gl. 2.4).

Das *HYBRID-Modell* von Yamamoto (1995) wurde zur Analyse von Speededness in High-Stakes-Tests entwickelt. Yamamoto (1995) nimmt an, dass Testteilnehmende, wenn sie unter Zeitdruck geraten, die verbleibenden Items durch ein zufälliges Antwortverhalten zu lösen versuchen. Dieses Antwortverhalten kann ein „Muster kreuzen“ (z. B. immer die erste Antwortoption), rein zufälliges Raten oder andere Muster bedeuten, die nicht mit einer (maximalen) Leistung in Zusammenhang stehen.

Damit ist dieses Antwortverhalten unabhängig von der Personenfähigkeit – somit zeigt sich auch zwischen Itemantworten am Testende und Personenfähigkeit kein Zusammenhang. In Gleichung 2.4 wird dies dadurch repräsentiert, dass die Itemdiskrimination auf Null fixiert wird. Die Personenvariable $\tilde{\theta}$ kann immer noch als Personenfähigkeit interpretiert werden, nur hat sie keinen Einfluss mehr auf die Itemantworten. Stattdessen wird für jedes Item ein Schwellenparameter angenommen, der dem Logit der Ratewahrscheinlichkeit entspricht. Wenn alle Items dieselbe Ratewahrscheinlichkeit haben, kann ein gemeinsamer Parameter geschätzt werden. Der Schwellenparameter muss nicht fixiert sein, sondern kann auch frei geschätzt werden.

Das HYBRID-Modell ist nicht auf die Analyse von Speededness beschränkt, sondern kann auch auf andere Testsituationen angewendet werden, zum Beispiel zur Analyse von Testbearbeitungsmotivation in Low-Stakes-Tests. Auch hier ist es plausibel, dass Testteilnehmende, wenn sie nicht mehr mit maximaler Leistung arbeiten, zu einem Rateverhalten übergehen (vgl. S. L. Wise & DeMars, 2005).

Das *Mixture-Rasch-Modell* von Bolt et al. (2002) nimmt an, dass es nur zwei latente Klassen gibt – eine für Personen ohne Leistungsabfall und eine für Personen mit Leistungsabfall. Da auch in diesem Modell die latenten Klassen mit spezifischen Schaltpunkten verknüpft sind, bedeutet dies, dass für alle Testteilnehmenden mit Leistungsabfall dieser an derselben Itemposition im Test einsetzt. Diese Itemposition kann a priori festgelegt werden, zum Beispiel aufgrund von Annahmen zum Testbearbeitungsverhalten, oder sie kann

aus den Daten geschätzt werden. Häufig wird als Itemposition $\frac{I}{2}$ verwendet, sodass der Leistungsabfall in der zweiten Testhälfte einsetzt (Bolt et al., 2002; Wollack et al., 2003)

Der Leistungsabfall wird im Modell von Bolt et al. (2002) über einen Anstieg in den Itemschwierigkeiten nach Beginn des Leistungsabfalls modelliert. Der Anstieg in den Itemschwierigkeiten ist itemspezifisch. Somit wird in erster Linie ein itemseitiger Positionseffekt modelliert. Auf der Seite der Testteilnehmenden wird ein binärer, personenseitiger Positionseffekt modelliert, da die Testteilnehmenden entweder einen Leistungsabfall zeigen oder den Test bis zum Ende mit maximaler Leistung bearbeiten.

Die Personenvariable entspricht auch nach dem Schaltpunkt der Personenfähigkeit, das heißt, der Leistungsabfall zeigt sich allein als Eigenschaft der Items. Die Itemparameter sind ebenfalls weiterhin als Itemdiskrimination und Itemschwierigkeit zu interpretieren, nur dass nach dem Schaltpunkt die Werte der Itemschwierigkeiten restringiert sind. Da im Mixture-Rasch-Modell nur zwei latente Klassen definiert sind, müssen die Klassenwahrscheinlichkeiten sowie die Verteilung der Personenfähigkeiten innerhalb der Klassen nicht restringiert werden, sondern können frei geschätzt werden.

Bolt et al. (2002) haben ihr Modell als Erweiterung zum 1PL entwickelt, sodass in der ursprünglichen Formulierung im Messmodell nur die Itemschwierigkeit vorkommt. Einige Anwendungen haben das Modell zum 2PL oder 3PL erweitert (z. B. Bolt, Mroch & Kim, 2003; Cao & Stokes, 2008; Suh et al., 2012) und unterscheiden sich darin, ob die zusätzlichen Itemparameter zwischen den latenten Klassen konstant gehalten oder separat geschätzt werden.

Das *Performance-Decline-Modell* von Jin und Wang (2014) nimmt wie das HYBRID-Modell an, dass Leistungsabfall an jeder Position im Test beginnen kann. Jin und Wang (2014) haben ihr Modell explizit für verschiedene Formen von Leistungsabfall (Speededness, Testbearbeitungsmotivation etc.) entwickelt. Anders jedoch als im HYBRID-Modell nehmen Jin und Wang (2014) an, dass sich der Leistungsabfall in einer Reduktion der

Personenfähigkeit zeigt. Dies wird durch einen Leistungsabfall-Parameter ausgedrückt, der von der Personenfähigkeit subtrahiert wird. Die Höhe dieses Parameters folgt einer linear fallenden Funktion in Abhängigkeit des Schaltpunkts. Dies bedeutet, dass der Leistungsabfall zu Beginn des Tests am größten ist und dann zum Testende sinkt. Die Personenvariable nach dem Schaltpunkt entspricht der Differenz von Personenfähigkeit und dem Leistungsabfall-Parameter. Die Itemparameter sind hingegen identisch mit denen vor dem Leistungsabfall.

Erweiterung zum Mehrgruppenmodell. Die drei Mischverteilungsmodelle können ferner so erweitert werden, dass die Analyse von Gruppenunterschieden im Leistungsabfall ermöglicht wird. Dazu werden die Modelle um eine manifeste (beobachtete) Gruppenvariable erweitert (vgl. Geiser, Lehmann & Eid, 2006) und je Gruppe wird ein separates Mischverteilungsmodell geschätzt. Während das Messmodell für die Itemantworten vor dem Schaltpunkt als konstant zwischen den Gruppen angenommen wird, können die mit dem Leistungsabfall assoziierten Parameter, das heißt Klassengrößen, Verteilung der Personenfähigkeiten, Schwellenparameter des Rateverhaltens (HYBRID-Modell), Itemschwierigkeiten nach dem Schaltpunkt (Mixture-Rasch-Modell) und Leistungsabfall-Parameter (Performance-Decline-Modell), gruppenspezifisch geschätzt werden.

Modellvergleiche. In verschiedenen Simulationsstudien wurden das Modell von Bolt et al. (2002) und das HYBRID-Modell von Yamamoto (1995) hinsichtlich der Reduktion des Bias in den Item- und Personenparametern untersucht. Suh, Kang, Wollack und Kim (2006) vergleichen die Schätzung der Personenfähigkeiten in einem Standard-IRT-Modell (d. h. ohne Berücksichtigung des Leistungsabfalls) mit dem Modell von Bolt et al. (2002) unter verschiedenen Kodierungen der Not-Reached-Items. Die Verzerrung in den Personenfähigkeiten ist geringer, wenn das Modell von Bolt et al. (2002) statt eines Standard-IRT-

Modells angewandt und wenn zusätzlich die Not-Reached-Items als fehlende Werte anstatt als falsche Antwort behandelt werden.

Die Auswirkungen auf die Schätzung der Itemparameter haben Suh et al. (2012) untersucht. Suh et al. (2012) vergleichen unter anderem das Modell von Bolt et al. (2002) mit dem HYBRID-Modell von Yamamoto (1995) hinsichtlich der Verbesserung der Itemparameterschätzung gegenüber einem Standard-IRT-Modell. Insgesamt zeigt sich, dass beide Modelle ähnlich gut geeignet sind, um die Verzerrung in den Itemschwierigkeiten durch den Leistungsabfall zu verringern. Die Diskriminationsparameter wurden im HYBRID-Modell (Yamamoto, 1995) überschätzt, während das Modell von Bolt et al. (2002) auch hier zu unverzerrten Schätzungen führte. Auch Suh et al. (2012) haben untersucht, ob sich die Behandlung der Not-Reached-Items auf die Parameterschätzung auswirkt. Es zeigten sich keine Unterschiede in der Reduktion des Bias, wenn Not-Reached-Items als fehlende Werte oder als falsche Itemantworten behandelt wurden.

In einer empirischen Studie vergleichen Bolt et al. (2003) die Schätzung der Itemparameter durch verschiedene Modelle, die für den Leistungsabfall am Testende kontrollieren. Bolt et al. (2003) finden, dass das Modell von Bolt et al. (2002) und das HYBRID-Modell (Yamamoto, 1995) die Verzerrung der Itemparameter am Testende ähnlich gut reduzieren können. Auch Cao und Stokes (2008) haben in einer empirischen Studie das Modell von Bolt et al. (2002) und das HYBRID-Modell (Yamamoto, 1995) verglichen, allerdings liegt in dieser Studie der Fokus auf der Modellpassung. Cao und Stokes (2008) finden, dass das HYBRID-Modell hinsichtlich Modellfit dem Modell von Bolt et al. (2002) überlegen ist.

Insgesamt zeigen die Befunde, dass Mischverteilungsmodelle in der Lage sind, die Verzerrung in den Item- und Personenparametern durch Effekte des Leistungsabfalls zu reduzieren. Die Befundlage ist allerdings zu klein, um generelle Aussagen über die Modellpassung in empirischen Studien zu treffen.

Kapitel 3

Fragestellungen und Ziele der Arbeit

3.1 Zusammenfassung des Forschungsstands

In nahezu allen LSAs erreichen nicht alle Testteilnehmenden das Testende, sodass es zu einer gewissen Anzahl von Not-Reached-Items kommt. Daneben zeigen sich ebenfalls in fast allen LSAs Positionseffekte. Positionseffekte bedeuten, dass die Lösungswahrscheinlichkeit geringer ausfällt, wenn das Item weiter hinten im Test platziert wird. Als personenseitiges Merkmal können Positionseffekte als Leistungsabfall betrachtet werden.

Beide Phänomene, Not-Reached-Items und Leistungsabfall, stellen eine Veränderung im Testbearbeitungsverhalten dar: Testteilnehmende zeigen nicht bis zum Testende ihre maximale Leistung, sondern brechen den Test an irgendeiner Itemposition ab oder führen den Test mit reduzierter Leistung zu Ende, die sich in der geringeren Lösungswahrscheinlichkeit für Items am Testende niederschlägt. Beide Phänomene können neben anderen Gründen durch eine geringe Testbearbeitungsmotivation oder durch Zeitdruck bedingt sein, wobei sich die Mechanismen zwischen Low-Stakes- und High-Stakes-Tests vermutlich unterscheiden.

Not-Reached-Items wie auch der Leistungsabfall stellen ein Problem für die Auswertung von Leistungstests in LSAs und die daraus abgeleiteten Ergebnisse dar. Die IRT-Modelle, die standardmäßig zur Auswertung von Leistungstests herangezogen werden, berücksichtigen keine Veränderung des Testbearbeitungsverhaltens. Beide Phänomene stellen deshalb eine Bedrohung für die Validität des Tests dar.

Bei Not-Reached-Items ergibt sich das Problem der Kodierung: Wie im vorherigen Kapitel 2 (Abschnitt 2.3, S. 20ff.) beschrieben wurde, implizieren die Kodierung als falsche Antwort und die Kodierung als fehlender Wert unterschiedliche Definitionen der Personenfähigkeit (vgl. Rohwer, 2013). Außerdem können beide Kodierweisen zu unterschiedlichen Fähigkeitsverteilungen und Gruppenunterschieden führen (vgl. Ludlow & O'Leary, 1999; Robitzsch, 2016). Ferner ist es plausibel, dass Not-Reached-Items zur MNAR-Kategorie

fehlender Werte gehören. Wird der Missingness-Prozess nicht berücksichtigt, kann dies zu einer Verzerrung in den Modellparametern und auch in den darauf basierenden Inferenzen (Gruppenunterschiede in der Personenfähigkeit, Zusammenhang von Personenfähigkeit mit weiteren Personenmerkmalen) führen.

Wird der Leistungsabfall im Test nicht berücksichtigt, kann dies ebenfalls zu Verzerrungen in den Itemparametern führen. Dies ist vor allem dann ein Problem, wenn die Itemparameter aus einer Erhebung und anderen Kontexten zum Linking oder zur Skalierung verwendetet werden sollen (s. Abschnitt 2.4, S. 31ff.). Da sich die Verzerrung auch in den Schätzungen der Personenfähigkeit niederschlagen kann (van der Linden, 2011; Mittelhaëuser et al., 2015a), kann die Nicht-Berücksichtigung des Leistungsabfalls auch zu verzerrten Ergebnissen hinsichtlich der Gruppenunterschiede in der Personenfähigkeit und der Leistungsentwicklung führen, besonders wenn das Ausmaß des Leistungsabfalls zwischen Gruppen oder Messzeitpunkten variiert (Mittelhaëuser et al., 2015a; G. Nagy, Retelsdorf et al., 2017).

Für beide Phänomene gibt es Hinweise, dass sie mit der Personenfähigkeit zusammenhängen und sich zwischen Gruppen (Schulform, Länder, Migrationshintergrund) unterscheiden. Da viele LSAs Fähigkeitsunterschiede zwischen Gruppen (z. B. Geschlecht, Migrationshintergrund, Schulform; vgl. Allmendinger et al., 2010; Dedering & Holtappels, 2010) untersuchen, sind differentielle Effekte von Not-Reached-Items und Leistungsabfall von besonderer Bedeutung, da sie die Validität der geschätzten Gruppenunterschiede bedrohen können.

3.1.1 Modellierung von Not-Reached-Items

Basierend auf Konzepten der *Missing-Data*-Forschung (Little & Rubin, 2002) wurden verschiedene Modelle vorgeschlagen, die den Zusammenhang von Not-Reached-Items und Personenfähigkeit modellieren. Die bisherigen Modelle lassen sich den Typen von Pattern-

Mixture-Modellen und Shared-Parameter-Modellen zuordnen. Pattern-Mixture-Modelle betrachten die Verteilung der Personenfähigkeit konditional zu verschiedenen distinkten Mustern (*Patterns*) von fehlenden Werten, Shared-Parameter-Modelle führen den Zusammenhang von fehlenden Werten und Personenfähigkeit auf Parameter, die die gemeinsame Verteilung beschreiben, zurück.

Einen weiteren Modelltyp bilden Selection-Modelle, die den Missingness-Prozess als abhängige Variable und Personenfähigkeit und weitere Kovariaten als deren Prädiktoren betrachten. Selection-Modelle eignen sich daher, um zu untersuchen, *wer* den Test *wann* abbricht. Bisher wurden jedoch keine Selection-Modelle für Not-Reached-Items vorgeschlagen.

Not-Reached-Items stellen einen Testabbruch dar und die Modellierung von Not-Reached-Items kann somit als Form einer Survivalanalyse für diskrete Messzeitpunkte betrachtet werden (vgl. Allison, 2014). Die Formulierung als Survivalanalyse impliziert ein Selection-Modell mit Testabbruch als abhängiger Variable und Personenfähigkeit als (zeitlich stabilem) Prädiktor. Die Itemposition entspricht der diskreten Messzeitpunktvariable. Die Formulierung als Survivalanalyse ermöglicht, die dort üblichen grafischen Werkzeuge zu verwenden, mithilfe derer auf einfache und anschauliche Art der Ausfallprozess im Testverlauf und komplexe Zusammenhänge mit den Prädiktoren dargestellt werden können.

Das bekannteste Modell für Not-Reached-Items stammt von Glas und Pimentel (2008). Dieses Modell wurde zwar als Shared-Parameter-Modell formuliert, es lässt sich aber als Selection-Modell reformulieren, wenn statt der Korrelation zwischen Personenfähigkeit und Missingness-Prozess eine Regression von Missingness-Prozess auf Personenfähigkeit modelliert wird. Das Modell von Glas und Pimentel (2008) weist daneben auch Ähnlichkeiten mit einer Survivalanalyse auf, weil der Missingness-Prozess anhand der Wahrscheinlichkeiten, den Test an einer bestimmten Itemposition abzubrechen, modelliert wird.

Als Survivalanalyse betrachtet ist das Modell von Glas und Pimentel (2008) sehr restriktiv, weil es davon ausgeht, dass die Wahrscheinlichkeit für einen Testabbruch monoton ansteigt und der Zusammenhang zwischen Personenfähigkeit und Testabbruch linear ist. Ferner wird für die Missingness-Variable, die den Testabbruch repräsentiert, eine Normalverteilung angenommen.

Da empirische Befunde für komplexe Beziehungen zwischen Testabbruch und Personenfähigkeit sprechen (z. B. G. P. Nagy, 1986), erscheint die Schätzung nichtlinearer Beziehungen zwischen Personenfähigkeit und Not-Reached-Items sinnvoll. Da Not-Reached-Items durch unterschiedliche Testbearbeitungsstrategien entstehen können, ist es auch plausibel, dass die Wahrscheinlichkeit für einen Testabbruch einen nichtmonotonen Verlauf zeigt und in Abhängigkeit von Personenfähigkeit und Kovariaten unterschiedlich verläuft (s. Dorans, Schmitt & Bleistein, 1992).

Zwar haben Rose et al. (2010) vorgeschlagen, eine nichtlineare Beziehung zwischen Personenfähigkeit und Not-Reached-Items anhand einer Polynom-Funktion zu schätzen, aber in diesem Ansatz ist der Testabbruch keine abhängige Variable, sondern ein Prädiktor für die Personenfähigkeit. Um Einblicke in den Mechanismus des Testabbruchs zu gewinnen, sind jedoch Modelle besser geeignet, die Testabbruch als die abhängige Variable betrachten. Ein solches Modell wurde bisher nicht entwickelt.

3.1.2 Modellierung von Leistungsabfall

Modelle für den Leistungsabfall nutzen entweder ein Multi-Matrix-Design für die Analyse von personenseitigen Positionseffekten oder verwenden Mischverteilungsmodelle, um latente Klassen von Personen mit einer Veränderung im Testbearbeitungsverhalten zu identifizieren. Der Vorteil des letztgenannten Modelltyps ist, dass diese Modelle auch eingesetzt werden können, wenn das Testdesign keine Analyse von Positionseffekten erlaubt.

Die Einsatzmöglichkeiten von Mischverteilungsmodellen zur Reduktion des Bias in den Item- und Personenparametern wurde in verschiedenen Studien dargestellt (Bolt et al., 2002; Suh et al., 2012). Zwei etablierte Mischverteilungsmodelle sind das Modell von Bolt et al. (2002) und das HYBRID-Modell von Yamamoto (1995). Vor kurzem wurde ein neues Modell von Jin und Wang (2014) vorgeschlagen. Bisher gibt es keine Studie, die das Modell von Jin und Wang (2014) mit den beiden anderen vergleicht; inwieweit das Modell von Jin und Wang (2014) eine bessere Alternative darstellt, ist somit noch eine offene Frage.

Die Untersuchung von Gruppenunterschieden im Leistungsabfall wurde bisher überwiegend mit dem Modell von Bolt et al. (2002) in Post-Hoc-Analysen untersucht. Dabei werden die Testteilnehmenden der latenten Klasse zugeordnet, für die sie die höchste Klassenwahrscheinlichkeit haben und die so gebildeten manifesten Klassen werden dann auf Unterschiede in der Zusammensetzung, zum Beispiel bezüglich des Geschlechterverhältnisses, untersucht (z. B. Bolt et al., 2002).

Ein Nachteil dieser Post-Hoc-Analyse ist, dass sie voraussetzt, dass die Testteilnehmenden mit großer Genauigkeit einer latenten Klasse zugeordnet werden können. Dies setzt eine hohe Separierbarkeit der latenten Klassen voraus. Besonders die Modelle, die für jede Itemposition eine latente Klasse definieren, weisen in der Regel aber eine Ungenauigkeit in der Klassenzuordnung auf. Letztendlich ist der exakte Schaltpunkt, an dem sich das Testbearbeitungsverhalten verändert, schwer zu schätzen. Alternativ kann ein Mehrgruppenmodell formuliert werden, dass den Vergleich der Verteilung der Parameter, die den Leistungsabfall beschreiben, zwischen den Gruppen ermöglicht. Eine Erweiterung zum Mehrgruppenmodell zur Analyse gruppenspezifischen Leistungsabfalls wurde bisher nur in einer empirischen Studie von Jin und Wang (2014) für ihr Modell verwendet, die beiden anderen Mischverteilungsmodelle (Bolt et al., 2002; Yamamoto, 1995) wurden bisher nicht zum Mehrgruppenmodell erweitert.

3.2 Fragestellungen der Dissertation

Anliegen dieser Dissertation ist es, bestehende Modellierungsansätze im Hinblick auf bestimmte Fragestellungen zu erweitern und miteinander zu vergleichen. Die Ziele sind dabei (1) die Einbeziehung von kategorialen Kovariaten, um Gruppenunterschiede in den Not-Reached-Items und im Leistungsabfall bei gleichzeitiger Kontrolle der Personenfähigkeit zu erfassen, und (2) die Lockerung von Restriktionen der bisherigen Modelle, die in bestimmten Anwendungen als unrealistische Annahmen betrachtet werden können.

Basierend auf dem Modell von Glas und Pimentel (2008) wird in der ersten Studie (Kap. 4, S. 53ff.) das *Mixture Discrete (Item) Sequence Event Model* (MDSEM) entwickelt. Wie auch im Modell von Glas und Pimentel (2008) wird über den Testabbruch, das heißt den Beginn der Not-Reached-Items, eine latente Missingness-Variable definiert. Das MDSEM betrachtet Missingness als abhängige Variable und entspricht damit einem Selection-Modell. Ähnlich wie im Modell von Köhler et al. (2015b) für ausgelassene Items modelliert auch das MDSEM die Missingness-Variable (für den Testabbruch) anhand diskreter Stützstellen, um eine nichtnormalverteilte Verteilung abbilden und nichtlineare Zusammenhänge zwischen latenter Personenfähigkeit und Testabbruch schätzen zu können. Darüber hinaus wird eine kategoriale Kovariate als weiterer Prädiktor für den Testabbruch in das Modell integriert und ermöglicht so die Schätzung differentieller Effekte.

In der zweiten Studie (Kapitel 5, S. 99ff.) werden drei Mischverteilungsmodelle (Bolt et al., 2002; Jin & Wang, 2014; Yamamoto, 1995) vertiefend analysiert. Als Alternative zur Post-Hoc-Analyse von Gruppenunterschieden im Leistungsabfall werden die drei Modelle zu Mehrgruppenmodellen erweitert, um Schulformunterschiede im Leistungsabfall untersuchen zu können. Außerdem wird analysiert, inwieweit die Berücksichtigung des differentiellen Ausmaßes von Leistungsabfall die Schätzung der Fähigkeitsunterschiede zwischen den Gruppen verändert.

In den folgenden beiden Kapiteln werden die Studien beschrieben. Diese Kapitel sind in englischer Sprache verfasst und – zumindest in Teilen – als Artikel in internationalen Fachzeitschriften publiziert. Die finalen Versionen der Artikel können aufgrund der editorischen Drucklegung von den Kapiteln der Dissertation abweichen. Durch die Formulierung als Zeitschriftenartikel enthalten die Kapitel zu den Studien jeweils auch einen Theorie- und Diskussionsteil und es ergeben sich gewisse Überlappungen mit den Kapiteln zur Theorie (Kap. 2, S. 15ff.) und Diskussion (Kap. 6, S. 139ff.) dieser Dissertation.

Kapitel 4

Studie 1:

Modellierung des Testabbruchs mittels Survivalanalysen

Die finale Version dieses Kapitels wurde in *Educational and Psychological Measurement* veröffentlicht, die Referenz lautet:

List, M. K., Köller, O. & Nagy, G. (2017). A semiparametric approach for modeling not-reached items. *Educational and Psychological Measurement*. Advanced online publication. Doi: 10.1177/0013164417749679

URL: http://journals.sagepub.com/doi/full/10.1177/0013164417749679

Die vorliegende Fassung des Artikels zur ersten Studie entspricht der letzten eingereichten Revision. Die finale Version des Manuskripts weicht davon in folgenden Hinsichten ab: (1) Aus Platzgründen wurde in der finalen Version auf das zweite empirische Beispiel (BIJU-Studie: Abschnitt Investigating the Onset of Not-Reached-Items in a Sample of 7th-Grade Students, S. 83ff.) verzichtet. (2) Darüber hinaus gibt es geringfügige Abweichungen durch die editorische Überarbeitung zur Drucklegung.

A Semiparametric Approach for Modeling Not-Reached Items

Abstract

Tests administered in studies of student achievement often have a certain amount of not-reached items (NRIs). The propensity for NRIs may depend on the proficiency measured by the test and on additional covariates. This article proposes a semiparametric model to study such relationships. Our model extends Glas and Pimentel's (2008) item response theory model for NRIs by (1) including a semiparametric representation of the distribution of the onset of NRIs, (2) modeling the relationships of NRIs with proficiency via a flexible multinomial logit regression, and (3) including additional covariates to predict NRIs. We show that Glas and Pimentel's (2008) and our model have close connections to event history analysis, thereby making it possible to apply tools developed in this context to the analysis of NRIs. Our model was applied to two timed low-stakes tests of mathematics achievement. Our model fitted the data better than Glas and Pimentel's (2008) model, and allowed for a more fine-grained assessment of the onset of NRIs. The results of a simulation study showed that our model accurately recovered the relationships of proficiency and covariates with the onset of NRIs, and reduced bias in the estimates of item parameters, proficiency distributions, and covariate effects on proficiency.

Key words: *educational assessment, item response theory, not-reached items, event history analysis, latent class analysis, nonlinear relations*

Introduction

Within scientific studies of student achievement, tests are typically administered with a time limit. As a consequence, students might not reach the end of a test within the allotted testing time, and this can lead to a special type of missing data, reflected in the number of not-reached items (NRIs). NRIs imply a monotone pattern of missing data; that is, all items located after the first item not reached are missing. Hence, the earlier the onset of NRIs, the more item responses are missing in a test. NRIs have received some attention in the psychometric literature because the onset of NRIs appears to be related to the proficiency measured by the test (Glas & Pimentel, 2008; Köhler et al., 2015a; Lawrence, 1993; Pohl et al., 2014). As a reaction to the problem, item response theory (IRT) models that make it possible to estimate the relationships between proficiency and the onset point of NRIs have been developed, with the approach suggested by Glas and Pimentel (2008) being most frequently used (e.g., Pohl et al., 2014). However, the model of Glas and Pimentel (2008), as well as many other approaches (e.g., Hutchison & Yeshanew, 2009), assumes a linear relationship between proficiency and the onset of NRIs: an assumption that might be questioned in many applications.

Linear relationships between NRIs and the proficiency measured by the test appear most plausible in testing situations in which NRIs can be conceived as pure reflections of test speededness (e.g., Evans & Reilly, 1972), such as in the case of high-stakes tests, where test takers invest full effort to respond to all items in the test. Here, students with higher proficiencies might solve the items at a higher pace, which means that they are more likely to reach the end of the test. However, even in situations in which test takers show their maximum performance, nonlinear relationships between proficiency and NRIs could exist because test takers at low proficiency levels might reach the end of the test by applying too simplistic or quick strategies to difficult items. In this case, test takers at an intermediate

proficiency level might then show the slowest solution behavior, leading to an earlier onset of NRIs. Relationships between proficiency and the onset point of NRIs appear likely to be complex in situations in which students are not motivated to show their maximum performance. In low-stakes assessments, test-taking behavior has been found to be related to test-taking motivation (S. L. Wise & DeMars, 2005), which means that NRIs could be impacted by motivational reactions to the test. Therefore, students with low proficiencies might either have an early onset of NRIs because they become frustrated with the test, or they might complete the test without investing much effort. Hence, the distribution of NRI onsets could be multimodal for certain levels of proficiency.

In addition, the amount of NRIs could also depend on person characteristics as well as on the proficiency being measured by the test (e.g., Dorans et al., 1992; Evans & Reilly, 1972; Köhler et al., 2015a; Schmitt et al., 1991). In principle, person variables could be related to the students' proficiencies and to the onset of NRIs. Hence, the question arises of whether NRIs can be fully predicted by the proficiency variable, or whether the person covariates have an additional impact on NRIs when proficiency is held constant. This question has some similarity with the concept of *differential test functioning* (Shealy & Stout, 1993) that refers to the question of whether a test works differently for examinees with the same proficiency but taken from different groups. Therefore, a finding that a covariate has an effect on NRIs while conditioning on proficiency indicates a *differential onset of NRIs* in examinees of the same proficiency, implying a systematic difference in the amount of information provided (i.e., the number of item responses preceding NRIs) to measure proficiency at the different levels of the covariate. Such differences could be due to different mechanisms, such as mastery in the test language or test-taking motivation, among others. Native speakers have been found to respond to more items in the allotted testing time (Schmitt & Bleistein, 1987; Sireci et al., 2008), and higher test-taking motivation has been found to be related to more time being spent on tasks (Scherer, Greiff & Hautamäki, 2015),

which could lead to an earlier onset of NRIs. However, as the background characteristics examined could be related to proficiency, in both examples, a rigorous test of the effects of students' background characteristics on NRIs requires the impact of proficiency on NRIs to be accounted for.

The aim of the present article is to provide a flexible and easy-to-use item response theory (IRT; Embretson & Reise, 2000) approach for modeling the onset of NRIs as a possibly nonlinear function of the proficiency measured by the test, as well as of additional person covariates. Our model combines a two-parameter logistic (2PL) IRT model (Birnbaum, 1968), applied to the item responses, with a latent class model (LCM; Formann, 1985), applied to the indicators of NRIs. Our LCM can be conceived as a semiparametric version of the continuous steps model for assessing the onset of NRIs suggested by Glas and Pimentel (2008). In our approach, the relationships of the onset of NRIs with the proficiency variable and the additional covariates were modeled via a multinomial logit regression, thereby allowing for nonlinear relationships. The newly proposed model was applied to two timed low-stakes tests of mathematics achievement; we thereby demonstrate its utility in applied settings. In a small simulation study, we further investigated whether the model correctly recovers the relationships of NRIs with proficiency and covariates, and whether our approach reduces biases in the estimates of item parameters, proficiency distributions, and covariate effects on proficiency that are often found in IRT models that disregard missing responses (e.g., Rose et al., 2010).

Relationships of NRIs with Proficiency and Covariates

Several studies have documented relationships between the onset of NRIs and the characteristics of test takers. Results suggest that the amount of NRIs is higher in ethnic minority groups (Dorans et al., 1992; Schmitt & Bleistein, 1987; Schmitt et al., 1991) but does not appear to differ between gender groups (Evans & Reilly, 1972; Schmitt et al., 1991; Wild,

Durso & Rubin, 1982). Some more recent studies have examined the relationships between NRIs and the proficiency measured by the test by adopting the IRT approach suggested by Glas and Pimentel (2008). These studies provide evidence for statistically significant relationships between proficiency and the onset of NRIs, but the pattern of results differed between tests and samples (Glas & Pimentel, 2008; Pohl et al., 2014). Most recently, Köhler et al. (2015a) studied the predictors of NRIs in reading tests implemented in several age groups. Their analyses revealed reading speed to be a strong and consistent predictor of NRIs, in that faster readers had a later onset of NRIs. Köhler et al. (2015a) employed the number of NRIs as a dependent variable in linear regression analyses. Similarly, researchers employing the model of Glas and Pimentel (2008) also did not investigate the nonlinear relationships that, as we have described above, appear plausible in the case of NRIs.

Investigating the effects of covariates on NRIs while simultaneously controlling for latent proficiency might be of interest for two reasons. First, in studies aiming to describe the distribution of student proficiencies in different subpopulations, the differential onset of NRIs indicates a threat to the validity of group comparisons as it means that groups differ in their test-taking behavior. Thus, group differences in proficiency might be different if respondents of the same proficiency level show the same test-taking behavior. Second, the effect of covariates on the onset of NRIs, while controlling for proficiency, could be a key research question in some applications. For example, researchers might hypothesize that a specific curricular intervention raises students' proficiencies and, in addition, enhances the pace at which students work on the test, thereby reducing the number of NRIs. To provide support for this hypothesis, a result indicating a differential onset of NRIs would be required.

Taken together, the differential onset of NRIs indicates that proficiency does not provide a sufficient explanation for differences in the onset of NRIs across different levels of a

covariate. Therefore, differential onsets of NRIs indicate that test takers that differ with respect to the covariate's value, but not to the level of proficiency, show different test-taking behavior. The differential onset of NRIs could indicate that the equivalence of measuring the proficiency variable across groups is violated. However, whether such a finding is considered as a threat to the validity clearly depends of the aims of the investigation.

IRT Models for Missing Responses and the Onsets of NRIs

Missing item responses in tests are regarded as problematic because they are likely to be related to the proficiencies being measured. As such, the missing data are nonignorable (MNAR; Little & Rubin, 2002), which means that missing data mechanisms need to be included in the model in order to prevent biased parameter estimates. To accomplish this task within the framework of the IRT, the full data likelihood that includes the vector of item scores $\boldsymbol{Y}$, a set of missing-data indicators $\boldsymbol{D}$, and possibly a set of covariates $\boldsymbol{X}$, that is $P(\boldsymbol{Y}, \boldsymbol{D}|\boldsymbol{X})$, needs to be considered. Consideration of $\boldsymbol{X}$ makes it possible to investigate whether the relationships between $\boldsymbol{Y}$ and $\boldsymbol{D}$ vanish once accounting for $\boldsymbol{X}$, so that the missing-data process is turned into a missing-at-random (MAR; Little & Rubin, 2002) process. Under MAR, $\boldsymbol{D}$ no longer contributes to the estimation of parameters that apply to $\boldsymbol{Y}$, and can therefore be ignored (e.g., Glas et al., 2015). Different types of modeling strategies have been applied to $P(\boldsymbol{Y}, \boldsymbol{D}|\boldsymbol{X})$. In typical IRT applications, the item scores $\boldsymbol{Y}$ are assumed to reflect one or multiple continuous proficiency variables, so that the relationships between $\boldsymbol{Y}$ and $\boldsymbol{D}$ are modeled via the relationships of the proficiency variables with $\boldsymbol{D}$ (e.g., Rose et al., 2015).

Pattern mixture IRT models (Little, 1994) aim to stratify the sample according to distinct missing-data patterns. They provide indications of MNAR patterns when the

proficiency variable (e.g., its mean) differs between strata. Practically, these models can be implemented either by means of multigroup IRT models, in which the groups are defined by distinct patterns of missingness, or by regressing the proficiency variables on indicators of the missing-data patterns, as well as on the covariates (e.g., Rose et al., 2017). In the context of NRIs, Rose et al. (2010) suggested regressing the continuous proficiency variable on the number of the individuals' NRIs. The model can be extended to include $\boldsymbol{X}$, as well as nonlinear relationships between proficiency and NRIs, for example, by using the polynomial functions of the amount of NRIs. Although the model is quite flexible and easy to use, it is not well suited for the purpose of studying the determinants of NRIs, because $\boldsymbol{D}$ is treated as an independent variable.

A second type of models assumes that additional latent variables underlie the missing-data indicators $\boldsymbol{D}$, which means that the relationships between proficiency and $\boldsymbol{D}$ are modeled via the relationships between latent variables. Most often, the joint distribution of latent variables is assumed to be multivariate and normal. These kinds of IRT models can be considered to belong to the family of *shared parameter models* (Wu & Carroll, 1988). They have been extended to multidimensional representations of $\boldsymbol{D}$, with the possibility of combining indicators of omissions with indicators of NRIs (Rose et al., 2017), and of including the covariates $\boldsymbol{X}$ affecting all latent variables (Glas et al., 2015). A drawback of these models is that the (conditional) multivariate normality assumption implies linear relationships between latent variables; this might be called into question. Koehler, Pohl, and Carstensen (2015b) relaxed the linearity assumption by formulating a two-dimensional IRT model for proficiency and omitted responses as a general diagnostic model (GDM; von Davier, 2008). GDMs allow any kind of multivariate distribution to be approximated by discretizing the latent variables into different prespecified levels. However, to the best of our knowledge, GDMs for missing responses have not yet been extended to include continuous covariates, and have not been applied to NRIs.

Selection models (Little & Rubin, 2002) refer to the third type of models that can be applied to account for MNAR patterns. Here, the full data likelihood $P(\boldsymbol{Y}, \boldsymbol{D}|\boldsymbol{X})$ is factorized into the distribution of $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$, $P(\boldsymbol{Y}|\boldsymbol{X})$, and the probability of missing data $\boldsymbol{D}$ given $\boldsymbol{Y}$ and $\boldsymbol{X}$, $P(\boldsymbol{D}|\boldsymbol{Y}, \boldsymbol{X})$, such that (e.g., Rose et al., 2017):

$$P(\boldsymbol{Y}, \boldsymbol{D}|\boldsymbol{X}) = P(\boldsymbol{Y}|\boldsymbol{X})P(\boldsymbol{D}|\boldsymbol{Y}, \boldsymbol{X}). \tag{4.1}$$

Because $\boldsymbol{D}$ is expressed as an outcome variable, selection IRT models provide a natural way for studying the determinants of missing responses, including NRIs. Within the IRT there are different ways of specifying models in which $\boldsymbol{D}$ depends on proficiency. In some models, it is assumed that the dependencies between all indicators $\boldsymbol{Y}$ and $\boldsymbol{D}$ can be fully explained by the proficiency variables. This assumption has been relaxed in other applications by including an additional continuous latent variable, so that $\boldsymbol{D}$ is simultaneously affected by multiple dimensions. In addition, the models can be extended to include covariates assumed to affect proficiencies, as well as the missing-data indicators. Furthermore, Bacci and Bartolucci (2015) relaxed distributional assumptions in these models by discretizing the latent variable. Their model is very flexible because it allows the relationships of proficiencies and covariates with missing responses to be item-specific. However, its drawback is that it includes many parameters that might not be reliably estimated in the case of a small percentage of missing values. Therefore, examining the determinants of the onsets of NRIs in the context of selection IRT models calls for an approach that consists of a suitable number of parameters to be estimated, and that allows for nonlinear effects of all the variables, including proficiencies. Furthermore, the model should allow for a flexible assessment of relationships.

NRIs reflect a special kind of missing data, because once an item response is missing in the sequence of test items, all responses that follow the first missing response are also

missing. Hence, this pattern of missing data can be regarded as being irreversible. Such situations are often at the core of longitudinal investigations that focus their attention on the risk (or hazard) that some irreversible events occur over time by employing methods known as event history analysis or survival analysis (Allison, 2014; Singer & Willett, 1993). Hence, the occurrence of NRIs over the sequence of test items, as represented by $P(\boldsymbol{D}|\boldsymbol{Y}, \boldsymbol{X})$ in Equation 4.1, can be examined by similar methods, with the difference that the (discrete) time points are replaced by discrete positions in a sequence of items. As we will show in the next section, when reformulated as a selection model, the model proposed by Glas and Pimentel (2008) can be regarded as a type of discrete-time event history model.

Glas and Pimentel's (2008) Model for NRIs in Speeded Tests

Glas and Pimentel (2008) considered NRIs and proposed a two-dimensional IRT model that includes a latent proficiency dimension and a second dimension indicating the number of items attempted by the examinees (i.e., a steps variable). The indicators of the proficiency variable are the actual item responses. In the case of dichotomous item responses, the proficiency variable is defined according to the 2PL model such that:

$$\text{logit}\, P(y_{ij} = 1|\theta_i) = \alpha_j(\theta_i - \beta_j) \ , \tag{4.2}$$

which means that the logit of the probability of a correct item response of individual i ($i = 1, 2, \ldots, N$) to item j ($j = 1, 2, \ldots, J$), y_{ij}, is a function of the individual's proficiency θ_i. In Equation 4.2, α_j and β_j stand for the discrimination and difficulty of item j.

The latent variable that assesses the onset of NRIs is measured by means of *response indicators* that are defined as follows: For each examinee, the vector of response indicators consists of a series of "1" for all items to which the examinee responds, followed by, at

most, one "0" for the first NRI, and *missing flags* for all subsequent not-reached items. For example, in a hypothetical 7-item test, an examinee who does not reach the last three items receives a vector of response indicators of $\boldsymbol{d}_i' = [1111099]$, where "9" indicates a missing value.

The steps variable underlying response indicators is defined by the *steps model* (Verhelst et al., 1997), which expresses the probability that a response is observed in a particular item position, given that all former item responses were observed. Glas and Pimentel (2008) presented applications in which the steps model included only a difficulty parameter, similar to the Rasch model (Rasch, 1960), such that

$$\text{logit}\, P(d_{ij} = 1|\xi_i) = \xi_i - \tau_j) \ , \tag{4.3}$$

where d_{ij} is the value of the response indicator for person i for item position j, ξ_i stands for the examinee i's steps variable with zero mean and unconstrained variance, and τ_j is a difficulty parameter of the response indicator in item position j. In the steps model, the difficulty parameters τ_j are constrained to follow a linear function across item positions, $\tau_j = t_0 + (j - J)\, t_1$, which means that only two parameters are estimated (i.e., t_0 and t_1). The latent variable ξ measures the number of steps taken by an examinee (i.e., number of items that are not NRIs): The lower the values of the steps variable ξ, the earlier the onset of NRIs.

The proficiency variable θ and the steps variable ξ are assumed to have a bivariate normal distribution with the correlation coefficient $\rho_{\xi,\theta}$. Positive correlations indicate that higher proficiencies are associated with later onsets of NRIs, whereas negative relationships indicate that higher proficiencies are related to earlier onsets of NRIs. In applications of Glas and Pimentel's (2008) model, the absolute size of the correlation coefficient is regarded as an indicator of MNAR missing-data patterns. However, as our focus is on the

relationship between proficiency and the onset of NRIs, we focused on a reparametrized version of their model in which ξ was treated as a dependent variable predicted by θ.

Relationships to Discrete-Time Event History Analysis

As previously mentioned, ξ can be treated as a variable depending on θ, such that

$$\xi_i = \gamma_{\xi,\theta}\theta_i + \zeta_{\xi,i} \ , \tag{4.4}$$

with $\gamma_{\xi,\theta}$ being a structural regression weight, and ζ_ξ standing for a normally distributed residual with zero mean and unconstrained variance that is uncorrelated with θ. By combining Equations 4.3 and 4.4, the relationships of θ with the response indicators d_j can be represented as

$$\text{logit}\left[P\left(d_{ij}=1|\theta_i,\zeta_{\xi,i}\right)\right] = \gamma_{\xi,\theta}\theta_i + \zeta_{\xi,i} - \tau_j \ , \tag{4.5}$$

showing that the model implies the same effect of the proficiency variable on the logit of each response indicator irrespective of its position.

Equation 4.5 has direct connections to discrete-time event history analysis, which models the effect of a variable on the probability that an irreversible event will occur, given that the event has not occurred before. If the ideas of event history analysis are applied to the phenomenon of NRIs, the focus is on $P\left(d_{ij}=0\right)$, marking the probability of not making the step from item j–1 to item j (given that all previous steps were taken), instead of on $P\left(d_{ij}=1\right)$, which refers to the probability of making the step from item j–1 to item j. To derive the hazard probability, the right hand side of Equation 4.5 could be multiplied by –1, such that

$$\text{logit}\left[P\left(d_{ij}=0|\theta_i,\zeta_{\xi,i}\right)\right] = \tau_j - \left(\gamma_{\xi,\theta}\theta_i + \zeta_{\xi,i}\right) \ . \tag{4.6}$$

Hence, the model of Glas and Pimentel (2008) can be reformulated as a kind of discrete-time event history model that is applied to the sequence of test items instead of to the sequence of time points, which means that the model can be understood as a *discrete (item) sequence event model* (DSEM). in Equation 4.6, θ serves as an explanatory variable, and the person variable ζ_ξ reflects a *frailty factor* (Allison, 2014; B. Muthén & Masyn, 2005) that accounts for heterogeneity in the hazards of NRI onsets not explained by θ. Similar to traditional discrete-time event models (Allison, 2014), the DSEM representation of the model of Glas and Pimentel (2008) assumes that the logits of all items' hazard probabilities are equally impacted by θ because $\gamma_{\xi,\theta}$ is not allowed to vary across items. In event history models, this assumption is known as the *proportional hazards assumption*, which means that the logit-hazard profiles (defined by $j = 1, 2, \ldots, J$) predicted by θ are proportional to one another (i.e., they have a common shape and are mutually parallel; Singer & Willett, 1993). Note that the DSEM presented in Equation 4.6 is more restrictive than conventional event history models because the parameters τ_j are constrained to follow a linear function, thereby forcing the baseline hazards of NRIs to increase over the course of the test. Typical discrete-time event models do not incorporate such assumptions and leave the τ-paramters unconstrained.

The DSEM formulation of Glas and Pimentel's model (2008) makes it possible to apply the graphical tools developed in the context of event history analysis to the onset of NRIs. Here, we focus on the *survival function*, which depicts the probability of "surviving" over a sequence of m ($m \leqq J$) items as a function of the explanatory variable θ. In the context of the present model, the survival function can be written as

$$\mathrm{P}\left(S_i \geq m|\theta_i, \zeta_{\xi,i}\right) = \prod_{j=1}^{m} P\left(d_{ij} = 1|\theta_i, \zeta_{\xi,i}\right) , \tag{4.7}$$

where $S = 1, 2, \ldots, J$ denotes the individual survival variable, whose value is equal to the last item a person has responded to. The survival function allows for a compact representation of the relationships between person variables and the probability of completing the test up to a specific point. In real applications, the survival function $P(S_i \geq m|\theta_i)$, defined over the full distribution of the frailty factor, might be more relevant than the survival function specified for specific combinations of θ and ζ_ξ (Equation 4.7). Deriving $P(S_i \geq m|\theta_i)$ requires determining the average of the function given in Equation 4.7 over the full distribution of ζ_ξ, which might turn out to be quite cumbersome in practice. This goal can be achieved by integrating over $\zeta_{\xi,i}$, or by simulating the distribution survival functions at the desired values of θ.

To sum up, the model suggested by Glas and Pimentel (2008) can be reformulated as a DSEM. This shows that Glas and Pimentel's (2008) model builds upon the proportional hazards assumption, that is, it specifies that θ has an equal impact on all response indicators. Their model also makes strong assumptions about the increasing baseline hazards (for a discussion, see Pohl et al., 2014). Furthermore, the model builds upon the assumption of a normally distributed frailty factor.

A Semiparametric Model for the Onset of NRIs

In this section we present a semiparametric version of the model of Glas and Pimentel (2008) which makes less strong assumptions about the distribution of the steps variable (i.e., the ξ-variable in Equation 4.4) and relaxes the proportional hazards assumption. In addition, we extend the model to include covariates, so that the model allows the differential onset of NRIs to be examined. Our approach builds upon the (continuous) 2PL model for item responses as represented in Equation 4.2, and on a semiparametric representation of the steps model given in Equation 4.4. In order to provide a metric for the steps variable that can be interpreted more easily, we propose a different parameterization of

Equation 4.3. We define a new steps variable δ that can be understood as a direct measure of the number of steps taken by an individual:

$$\text{logit}\left[P\left(d_{ij}=1|\delta_i\right)\right]=\lambda\left(\delta_i-j\right). \tag{4.8}$$

in Equation 4.8, λ stands for a discrimination parameter that applies to all response indicators. The difficulty parameters given in Equation 4.3 are replaced by the item indexes $j = 1, 2, \ldots, J$. Because the difficulties in Equation 4.8 are fixed, the mean of δ can now be estimated. The latent variable δ is measured in units of item positions, such that $\delta_i = j$ provides a probability of .5 of providing an item response in position j. The probabilities of providing responses to items preceding j are higher than .5 and the probabilities for item positions after j are lower than .5. Because the items are assumed to be equally spaced, the probability curve follows a logistic function. The steepness of the function is driven by λ, with higher values indicating more abrupt changes in the probability of responding around the individual inflection point δ_i.

In our semiparametric version of the steps model we specify δ to have a discrete distribution, with K support points δ_k ($k = 1, 2, \ldots, K$) that are freely estimated, in order to derive a flexible representation of the distribution δ-variable. This approach is equivalent to a *located latent class model* (Formann, 1985), which includes a latent class variable c with K categories and class proportions π_k with $0 \leqq \pi_k \leqq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. The purpose of our LCM application was to relax the assumption of a normally distributed steps variable (and, therefore, of a normally distributed frailty factor), and to approximate the unknown distribution of δ (Masyn, 2009; B. Muthén & Masyn, 2005). Our approach makes it possible to relax the linearity assumption included in the model of Glas and Pimentel (2008), which implies the proportional hazards assumption. Our approach can be considered as an *indirect application* of a LCM (Bauer, 2005), where the primary task is not to classify

individuals, but to relax distributional and functional assumptions. The LCM part can be represented by altering Equation 4.8 to

$$\text{logit}\left[P\left(d_{ij}=1|c_i=k\right)\right]=\lambda\left(\delta_k-j\right) \tag{4.9}$$

As shown in Equation 4.9, δ is assumed to be constant within latent classes, which means that, for each class, only one pattern of probabilities exists for all J response indicators. The values of δ_k can be considered as the *support points* of a nonparametric approximation of an arbitrary continuous distribution, whereas the latent class proportions π_k can be considered as the weights associated with support points (Aitkin, 1999; Bacci & Bartolucci, 2015).

The model for response indicators has connections to mixture discrete-time event history analysis (B. Muthén & Masyn, 2005). As the pattern of probabilities of response indicators is constant within each latent class, each class provides one survival function:

$$P\left(S_i\geq m|c_i=k\right)=\prod_{j=1}^{m}P\left(d_{ij}=1|c_i=k\right)\ . \tag{4.10}$$

Furthermore, the marginal survival function can be derived by summing over the class-specific survival functions (Equation 4.10) weighted by their class proportions:

$$P\left(S_i\geq m\right)=\sum_{k=1}^{K}\pi_k P\left(S_i\geq m|c=k\right)\ . \tag{4.11}$$

The semiparametric approach allows for heterogeneity in the latent steps variable δ, since the LCM approach circumvents the assumption of a normally distributed latent steps variable.

The main motivation behind our semiparametric approach was to relax the proportional hazards assumption implicitly made in the model of Glas and Pimentel (2008; see

Equation 4.6). Here, we estimate the impact of θ on δ by means of multinomial regression, relating the latent class variable c to θ, such that

$$P\left(c_i = l|\theta_i\right) = \frac{\exp\left(\nu_l + \gamma_{l,\theta}\theta_i\right)}{\sum_{k=1}^{K}\exp\left(\nu_k + \gamma_{k,\theta}\theta_i\right)} \ , \tag{4.12}$$

where ν_l is the structural intercept specific to class $c = l$, and $\gamma_{l,\theta}$ is the regression weight relating the proficiency variable θ to the latent class $c = l$. We impose the typical identification restriction of fixing the multinomial parameters of the class $c = K$ to zero.

By using the multinomial regression (Equation 4.12), the proportional hazards assumption is relaxed. Since, for each latent class c, a regression coefficient is estimated, an overall nonlinear relationship between proficiency and the onset of NRIs is modeled. For example, the analysis could uncover that θ is only related to membership in classes representing later onsets of NRIs, whereas the continuous approach specifies that θ shows an equal impact on all response indicators. Alternatively, findings could indicate that low levels of proficiency are simultaneously related to membership in classes representing an early onset of NRIs and classes representing a late onset of NRIs.

The model described in Equations 4.9 and 4.12 fits into the framework of mixture discrete-time event history models (B. Muthén & Masyn, 2005). Therefore, we refer to this model as *mixture discrete (item) sequence event model* (MDSEM). In the MDSEM the survival function that depends on θ is derived by summing over the class-specific survival functions (Equation 4.10) weighted by the class probability predicted by θ (Equation 4.12):

$$P\left(S_i \geq m|\theta_i\right) = \sum_{k=1}^{K} P\left(c_i = k|\theta_i\right)P\left(S_i \geq m|c_i = k\right) \ . \tag{4.13}$$

The MDSEM provides a flexible and easy-to-use framework for deriving the survival functions at fixed values of θ. Deriving the survival functions for fixed levels of θ no

longer requires integration over a continuous frailty factor, as is the case in the DSEM (Equation 4.7).

Introducing Additional Covariates

We now address the case of introducing covariates into the model. In order to keep the presentation simple, we focus on the case with a single covariate x, but note that multiple covariates could be included simultaneously. We begin by using x for predicting θ. Here, we assume a linear relationship, such that

$$\theta_i = \kappa_\theta + \gamma_{\theta,x} x_i + \zeta_{\theta,i} \tag{4.14}$$

where κ_θ is a structural intercept, $\gamma_{\theta,x}$ is the regression weight of x predicting θ, and ζ_θ is a normally distributed residual term with an expectation value of zero that is assumed to be uncorrelated with all other variables in the model.

The relationship between latent classes and predictors θ and x is now given as

$$P\left(c_i = l|\theta_i, x_i\right) = \frac{\exp\left(\nu_l + \gamma_{l,\theta}\theta_i + \gamma_{l,x}x_i\right)}{\sum_{k=1}^{K} \exp\left(\nu_k + \gamma_{k,\theta}\theta_i + \gamma_{k,x}x_i\right)} \ , \tag{4.15}$$

where the parameters are as defined as before (Equation 4.12) and subjected to the same identification constraints. The only difference is that the multinomial regression part is extended by the covariate x and its class-specific multinomial regression weight $\gamma_{l,x}$.

Equation 4.15 is of key importance to the suggested model. More specifically, the hypothesis that missing data caused by NRIs do not depend on the covariate x (or on a set of covariates $\boldsymbol{X}$) implies that $\gamma_{1,x} = \gamma_{2,x} = \ldots = \gamma_{K,x} = 0$, assuming that the covariate's effects on class membership are transmitted via its impact on the proficiency variable (Equation 4.14). When the model is estimated by maximum likelihood, this hypothesis can be evaluated by means of a likelihood ratio test (LRT) that compares the data likelihood

of a full model with a nested model in which the relationships with the latent class variable c are accessed via Equation 4.12, that is, by setting $\gamma_{1,x} = \gamma_{2,x} = \ldots = \gamma_{K,x} = 0$. A statistically significant LRT provides evidence for the differential onset of NRIs, depending on the covariate considered.

If the covariate x is found to predict latent class membership, its impact can be visualized by using the survival function, evaluated with selected combinations of θ and x:

$$P\left(S_i \geq m|\theta_i, x_i\right) = \sum_{k=1}^{K} P\left(c_i = k|\theta_i, x_i\right) P\left(S_i \geq m|c_i = k\right) \ . \tag{4.16}$$

Model Estimation and Implementation

The MDSEM can be estimated by maximum likelihood estimation. Indeed, Guo, Wall, and Amemiya (2006) have outlined the estimation of a general class of models, of which the MDSEM is a special case. The joint distribution of the item responses $\boldsymbol{Y}$ and NRI indicators $\boldsymbol{D}$ can be written as (by dropping the symbolic representation of model parameters)

$$P\left(\mathbf{Y}, \mathbf{D}|\mathbf{X}\right) = \sum_{k=1}^{K} \int P\left(\mathbf{D}|c_i = k\right) P\left(\mathbf{Y}|\theta\right) P\left(c_i = k|\theta, \mathbf{X}\right) P\left(\theta|\mathbf{X}\right) \ d\theta, \tag{4.17}$$

with the full data likelihood function L given by

$$L = \prod_{i=1}^{N} P\left(\mathbf{Y_i}, \mathbf{D_i}|\mathbf{X_i}\right) = \prod_{i=1}^{N} \int P\left(\mathbf{Y_i}, \mathbf{D_i}, c_i, \theta_i|\mathbf{X_i}\right) d\left(c_i, \theta_i\right), \tag{4.18}$$

whereby the integral includes the continuous integral over θ, as well as summation over c.

Guo, Wall, and Amemiya (2006) have shown that the model parameters can be estimated by means of the expectation maximization (EM) algorithm, as well as by a Gaussian quadrature with a quasi-Newton algorithm. Hence, the MDSEM can be estimated with different computer programs, including Latent Gold (Vermunt & Magidson, 2005) and Mplus (L. K. Muthén & Muthén, 1998-2012). In the present article, we used Mplus, which combines the aforementioned algorithms to a so-called accelerated EM algorithm (EMA). Model estimation starts with the EM algorithm but changes to the quasi-Newton algorithm if EM becomes slow.

There are several issues that need to be considered in practice. The first issue is how to define the metric of the proficiency variable. In most IRT applications, this issue is resolved by standardizing the distribution of θ (i.e., $M = 0$, $SD = 1$). Since, in the general case, θ is specified as an endogenous variable that is impacted by covariates, we suggest freely estimating the (residual) variance term, fixing the mean or the structural intercept of θ to zero (Equation 4.14), and constraining the discrimination parameters such that they are, on average, one. The latter constraint allows the (residual) variance of the proficiency variable to be freely estimated and does not alter its units of measurement when including covariates to predict θ.

The second issue is that NRIs might not exist in the first positions of a test. As a consequence, all response indicators gathered before the first onset of NRIs have a constant value across all respondents, which means that they should be disregarded in the process of model estimation.

The third issue pertains to the optimal number of latent classes. In applications of latent class analysis, the decision concerning the number of classes to use is typically based on measures of goodness-of-fit, such as the Bayesian information criterion (BIC; Schwarz, 1978; Nylund et al., 2007). Users need to be aware that the optimal number of latent classes could also depend on the covariates used for predicting latent class membership

(Lubke & Muthén, 2005). We suggest basing the decision about the number of classes K on the full MDSEM and keeping K constant across different versions of the model (e.g., models where the covariates are excluded) to make sure that results are not affected by the use of different numbers of classes. Some researchers have suggested identifying the number of latent classes prior to the inclusion of covariates (Kim, Vermunt, Bakk, Jaki & Horn, 2016; Nylund-Gibson & Masyn, 2016). This approach is useful in *direct applications* of the LCM (Bauer, 2005) that require the categorical latent variable to exist independent of the covariates included because individuals' class membership is substantively interpreted. However, the MDSEM is based on an indirect application of the LCM that does not aim to categorize individuals, but only to relax parametric assumptions.

Finally, one problem with maximum likelihood estimation for mixture IRT models is that the solution can converge to a local rather than the global maximum (Finch & French, 2012). Therefore, the usage of multiple random starting values is recommended to ensure replication of the best likelihood value (Lubke & Muthén, 2005).

Empirical Illustrations

In the next sections we report on our application of the proposed MDSEM to two different datasets of timed low-stakes mathematics tests taken from typical large-scale studies. One sample consisted of apprentices, the other one of students in the seventh grade. These applications served three purposes. First, we compared the MDSEM to the model suggested by Glas and Pimentel (2008) thereby demonstrating the flexibility gained by implementing their model in a semiparametric framework. Second, we exemplify how to use the MDSEM to evaluate a test for the differential onset of NRIs while holding the proficiency variable constant. Third, we exemplify the use of graphical procedures to aid the interpretation of model results, while focusing on the survival function.

The models were implemented in Mplus 7.4 (L. K. Muthén & Muthén, 1998-2012) by using the EMA algorithm using standard integration with 15 integration points for the proficiency variable. All models were estimated using multiple sets of random starts. In all cases, the best log-likelihood was replicated. In order to determine the number of classes, we estimated a series of models ranging from 3 to 8 latent classes. The decision concerning the number of classes was based on the BIC.

Investigating the Onset of NRIs in a Sample of Apprentices

Sample and Procedure

The sample was taken from the study "Mathematics and Science Competence in Vocational Education and Training" (ManKobE; e.g., Retelsdorf, Lindner, Nickolaus, Winther & Köller, 2013). It encompassed apprentices in their first year of vocational education and training (VET) in mathematics and science-related occupations, namely, industrial clerks and different technical professions (e.g. industrial and laboratory technicians; further referred to as *technicians*). The test was designed to assess mathematical skills at the core of VET for industrial clerks. The test contained only tasks that could, in principle, be solved with the mathematical knowledge acquired in regular schooling, but the problems presented were embedded in an organizational context typical for industrial clerks. Hence, in this analysis, we expected that industrial clerks would have higher proficiency, and we hypothesized that they would show a later onset of NRIs than technicians because the context in which the items were presented was more familiar to clerks.

We considered the data of $N = 967$ apprentices at the beginning of their VET (average time in VET of about 3 months); cases with less than three valid responses in the whole test and those with missing information on the covariate considered were excluded. From all test takers, $n = 214$ cases were in VET for industrial clerks; the remaining apprentices

were in VET for technicians ($n = 753$). On average, apprentices were 18.70 years old ($SD = 2.88$). The test considered consists of 20 dichotomously scored items and it was administered with a time limit of approximately 15 minutes. NRIs were first observed in item position $j = 5$. Therefore, response indicators for the first four items were not included in the analysis.

Results

Descriptive Results. Only 28 % of the sample completed the first three quarters of the test and only 16 % reached the last item. The sample-based baseline hazard function of the onsets of NRIs is depicted in Figure 4.1a. It appeared that the hazard rates for NRIs did not constantly increase across item positions but rather reached a maximum after the first three quarters of the test (i.e., in position 15). In addition, the hazard function given in Figure 4.1a appeared to be constituted of several peaks. This led us to expect that the semiparametric steps model was likely to identify several latent classes that are sharply separated from each other.

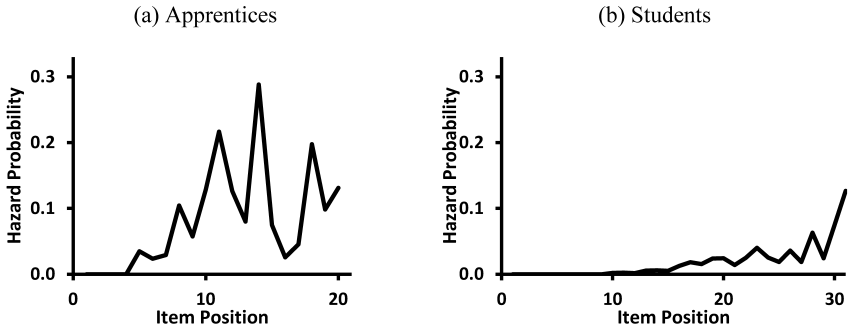

Figure 4.1: Sample-estimated hazard probabilities of onsets of NRIs.

Nonparametric Representation of the Distribution of Steps Variable. As shown in Table 4.1, the model with $K = 6$ classes achieved the best fit in terms of the BIC, and we therefore decided on six classes. In this model, the discrimination parameter of response

Table 4.1: Fit Statistics for MDSEM with Different Numbers of Latent Classes (K)

		No. Par.	Log-Likelihood	AIC	BIC
Apprentices	K = 3	51	-7,066	14,233	14,482
	K = 4	55	-7,056	14,222	14,490
	K = 5	59	-7,021	14,159	14,447
	K = 6	63	-6,969	14,064	14,371
	K = 7	67	-6,963	14,060	14,387
	K = 8	71	-6,954	14,050	14,396
Students	K = 3	73	-91,501	183,148	183,632
	K = 4	77	-91,476	183,106	183,617
	K = 5	81	-91,462	183,086	183,622
	K = 6	85	-91,450	183,069	183,633
	K = 7	89	-91,454	183,086	183,676
	K = 8	93	-91,448	183,082	183,699

indicators was estimated to be $\hat{\lambda} = 2.16$ ($SE = 0.116$). Hence, a relatively sharp step in the probability of responding to items around the inflection points (support points) was estimated. This pattern was expected, on the basis of the sharply peaked hazard function (Figure 4.1a). The estimates of support points and posterior class proportions are summarized in Table 4.2. As shown there, the support points were quite evenly distributed, and were close to the peaks of the empirical hazard function (Figure 4.1a). As indicated by the estimated posterior proportions, the distribution of the steps variable appeared to have two modes. For $c = 6$, the support point was somewhat above the maximum test length of 20 items (i.e., $\hat{\delta}_6 = 20.92$) and this class received a proportion of $\hat{\pi}_6 = .19$. This class describes test takers who were likely to complete the test. Class $c = 3$ received the largest proportion, indicating that around 30 % of the test takers ($\hat{\pi}_3 = .32$) switched to NRIs around the middle position of the test ($\hat{\delta}_3 = 10.43$).

Comparison to the Model of Glas and Pimentel (2008). We now turn to the comparison of the MDSEM, in which the indicator of group membership was discarded, with the model for NRIs proposed by Glas and Pimentel (2008). The first step was to estimate the MDSEM without considering group membership and to estimate Glas and

Table 4.2: Proportions and Support Points of the Nonparametric Representation of the Distribution of the Steps Variable

		$\hat{\pi}_k$	$\hat{\delta}_k$	(SE)
Apprentices	k = 1	.05	4.75	(0.167)
	k = 2	.16	7.65	(0.078)
	k = 3	.32	10.43	(0.052)
	k = 4	.19	13.48	(0.064)
	k = 5	.09	17.62	(0.099)
	k = 6	.19	20.92	(0.123)
Students	k = 1	.08	19.57	(0.320)
	k = 2	.14	24.91	(0.249)
	k = 3	.31	33.95	(0.298)
	k = 4	.48	36.89	(0.392)

Pimentel's (2008) model. The MDSEM contained 57 free parameters and achieved a log-likelihood value of LL = –6,994 (AIC = 14,103, BIC = 14,381). The model of Glas and Pimentel (2008) contained fewer parameters (44) and achieved a lower log-likelihood value of LL = –7,172 (AIC = 14,432, BIC = 14,647). In addition, the information indices were clearly in favor of the MDSEM.

Despite the difference in model-data fit, both models provided nearly identical estimates for the measurement part of the proficiency variable. The estimates of item discriminations showed only minor deviations between the models [mean absolute deviation (MAD) = 0.030], and the same was true for item difficulties (MAD = 0.080). Furthermore, the variance of the latent proficiency variable was estimated by the model of Glas and Pimentel (2008) as $\hat{\sigma}^2_\theta = 1.95$ ($SE = 0.257$), and by the MDSEM as $\hat{\sigma}^2_\theta = 1.91$ ($SE = 0.248$). Marked differences were found for the estimated relationship between θ and the steps variable. In the model of Glas and Pimentel (2008), the unstandardized regression weight was estimated to be $\hat{\gamma}_{\xi,\theta} = -0.21$ ($SE = 0.044$, $p \leq .001$) and a standardized counterpart to be $\hat{\gamma}^{\text{stnd}}_{\xi,\theta} = -.39$ ($SE = 0.100$, $p \leq .001$). This result documents a relationship of medium strengths that indicates that higher levels of proficiency were related to earlier onsets of NRIs.

The multinomial logit coefficients determined by the MDSEM are reported in Table 4.3. The intercept parameters mirrored the latent class proportions (Table 2). The regression weights represent the change in the log-odds of belonging to class $c = l$ relative to the reference class $c = 6$ for one unit increase in θ. The analyses uncovered a pattern reflecting a curvilinear relationship and indicating that the chance of being classified into classes 2 to 4, as opposed to class 6, increased with higher values of θ. To gain a better insight into the relationships, Table 3 also reports the class probabilities expected for values of θ at the 10th, 50th, and 90th percentiles of the (normal) proficiency distribution. As can be seen, students of low ability tended to be more evenly distributed across latent classes, whereas high ability students became more concentrated in the interim classes, especially in class 3.

Table 4.3: Multinomial Logit Coefficients of the Regression of Latent Class on the Proficiency Variable (θ), and Predicted Class Probabilities (PCP) as Selected Values of θ (10th, 50th, and 90th Percentiles)

		Multinomial Logit Coefficients				PCP		
		$\hat{\nu}_k$	(SE)	$\hat{\gamma}_{k,\theta}$	(SE)	θ=-1.77	θ=0.00	θ=1.77
Apprentices	k = 1	-1.25	(0.196)***	0.08	(0.184)	.08	.05	.03
	k = 2	-0.05	(0.123)	0.38	(0.113)**	.15	.17	.17
	k = 3	0.59	(0.107)***	0.49	(0.098)**	.23	.32	.40
	k = 4	0.11	(0.117)	0.46	(0.104)**	.15	.20	.24
	k = 5	-0.69	(0.148)***	0.20	(0.124)	.11	.09	.07
	k = 6	–	–	–	–	.30	.18	.09
Students	k = 1	-1.51	(0.207)***	1.65	(0.231)***	.03	.09	.09
	k = 2	-1.11	(0.187)***	2.19	(0.251)***	.02	.14	.26
	k = 3	-0.23	(0.358)	2.15	(0.240)***	.06	.34	.59
	k = 4	–	–	–	–	.89	.43	.06

Note. ** p <.01. *** p < .001.

To facilitate a better comparison of the predictions made by the two models, Figure 4.2 provides the survival curves derived at the 10th, 50th, and 90th percentiles of the (normal) proficiency distribution. The survival functions were markedly different. More specifically, Glas and Pimentel's (2008) model predicted that the survival curves were already different

at the onset of the first NRIs (i.e., starting from $j = 5$). In contrast, the MDSEM revealed that the onset of NRIs, and hence the survival curves, started to be impacted by proficiency from the middle position (around $j = 10$) on, which means that the occurrence of NRIs in the second quarter of the test (between the 5th and 10th item position) was not related to proficiency. In addition, the MDSEM indicated larger differences in the survival probabilities in the last quarter of the test, compared to the model of Glas and Pimentel (2008). Moreover, the survival curves provided by the MDSEM were not as smooth as the curves provided by the model of Glas and Pimentel (2008), which were close to a linear function. The survival curves of the MDSEM appeared to reflect the peaked nature of the hazard functions (Figure 4.1a).

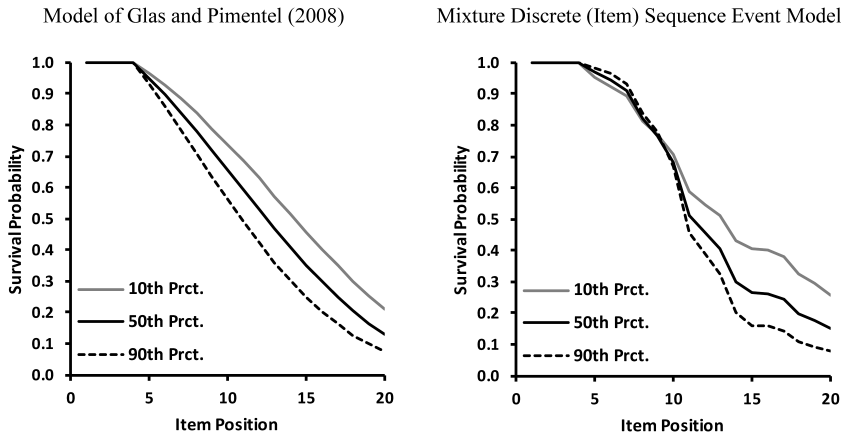

Figure 4.2: Survival functions determined for the 10th, 50th, and 90th percentile (Prct.) of the proficiency distribution determined on basis of the model of Glas and Pimentel (2008) and the MDSEM in the sample of apprentices.

Differential Onset of NRIs. The last issue considered here concerns the question of whether membership in the two VET fields was related to the onset of NRIs. Hence, we now turn to the full MDSEM, in which the group membership was included (0 = *technicians*, 1 = *clerks*). The goodness-of-fit of the full MDSEM is reported in Table 4.1. In order to find out whether group membership was related to the onset of NRIs over and above proficiency,

we estimated a more constrained version of the MDSEM, in which the multinomial weights of group membership were fixed to zero. The fit of the models was compared via an LRT and provided a statistically significant result, $\chi^2(df = 5) = 24.9$ ($p \leq .001$), indicating that group membership was related to the onset of NRIs over and above the proficiency variable.

The analysis provided the expected results. Clerks were found to have a higher proficiency level ($\hat{\gamma}_{\theta,x} = 0.66$, $SE = 0.130$, $p \leq .001$), and a lower probability of belonging to the classes $c = 2$ and $c = 3$ than to the reference class $c = 6$, compared to the technicians. These results are shown by the multinomial regression weights provided in Table 4.4.

The regression weights for the proficiency variable predicting class membership were very similar to the results provided in Table 4.3. Table 4.4 also reports the class probabilities predicted by proficiency and group membership. The corresponding probabilities were evaluated at the 10th, 50th, and 90th percentiles of the combined proficiency distribution with equally weighted groups. Clerks of the same proficiency level were more likely to belong to the latent classes associated with a later onset of NRIs. This finding is visualized by the survival functions in Figure 4.3a. The survival function already differed between groups right after the first onset of NRIs (at $j = 5$). In this region, survival did not depend on proficiency. The most pronounced group differences were determined for the third quarter of the test, where the survival function showed a steeper decrease at all levels of proficiency in the group of technicians. In the fourth quarter of the test, the survival curve was flatter for technicians, but the survival probability was still lower compared to the group of industrial clerks.

Summary

With this application, we intended to provide an example for an application of the proposed MDSEM in a low-stakes test characterized by a high prevalence of NRIs. As we have shown,

Table 4.4: Multinomial Logit Coefficients of the Regression of Latent Class on the Proficiency Variable (θ), and Predicted Class Probabilities (PCP) as Selected Values of θ (10th, 50th, and 90th Percentiles in the Combined Sample) for subgroups

		$\hat{\nu}_k$	(SE)	$\hat{\gamma}_{k,\theta}$	(SE)	$\hat{\gamma}_{k,x}$	(SE)
Apprentices	k=1	-1.10	(0.206)***	0.14	(0.190)	-0.80	(0.523)
	k=2	0.05	(0.138)	0.42	(0.118)**	-0.67	(0.305)*
	k=3	0.67	(0.122)***	0.54	(0.103)**	-0.74	(0.257)**
	k=4	0.10	(0.135)	0.48	(0.109)**	-0.21	(0.266)
	k=5	-0.88	(0.181)***	0.15	(0.130)	0.52	(0.315)
	k=6	–	–	–	–	–	–
Students	k=1	-2.09	(0.164)***	2.26	(0.310)***	-1.49	(0.287)***
	k=2	-2.02	(0.170)***	2.73	(0.333)***	-1.28	(0.267)***
	k=3	-1.24	(0.385)**	2.73	(0.337)***	-1.24	(0.342)***
	k=4	–	–	–	–	–	–

		PCP: Technicians			PCP: Industrial Clerks		
		θ=-1.45	θ=0.33	θ=2.12	θ=-1.45	θ=0.33	θ=2.12
Apprentices	k=1	.08	.05	.03	.04	.03	.02
	k=2	.16	.18	.19	.10	.13	.15
	k=3	.25	.35	.45	.15	.24	.33
	k=4	.15	.20	.22	.16	.22	.28
	k=5	.09	.07	.04	.20	.16	.11
	k=6	.28	.15	.07	.35	.22	.11
		PCP: Nonacademic track			PCP: Academic track		
		θ=-0.61	θ=0.56	θ=1.73	θ=-0.61	θ=0.56	θ=1.73
Students	k=1	.03	.13	.11	.01	.06	.09
	k=2	.02	.18	.27	.01	.10	.26
	k=3	.05	.39	.60	.02	.23	.59
	k=4	.90	.29	.02	.97	.60	.06

Note. * p $<$ .05. ** p $<$.01. *** p $<$ 0.001.

the MDSEM provides a method for detecting nonlinear patterns of the onset points of NRIs by using a semiparametric parameterization of the steps model. Compared to the parametric NRI model provided by Glas and Pimentel (2008), the MDSEM allows for a more flexible representation of the test survival function. In the present case, the MDSEM provided a survival curve that better reflected the peaked nature of the hazard function. In addition, the MDSEM does not rely on the proportional hazard assumption and was therefore able to identify regions where the onset of NRIs depended on person variables,

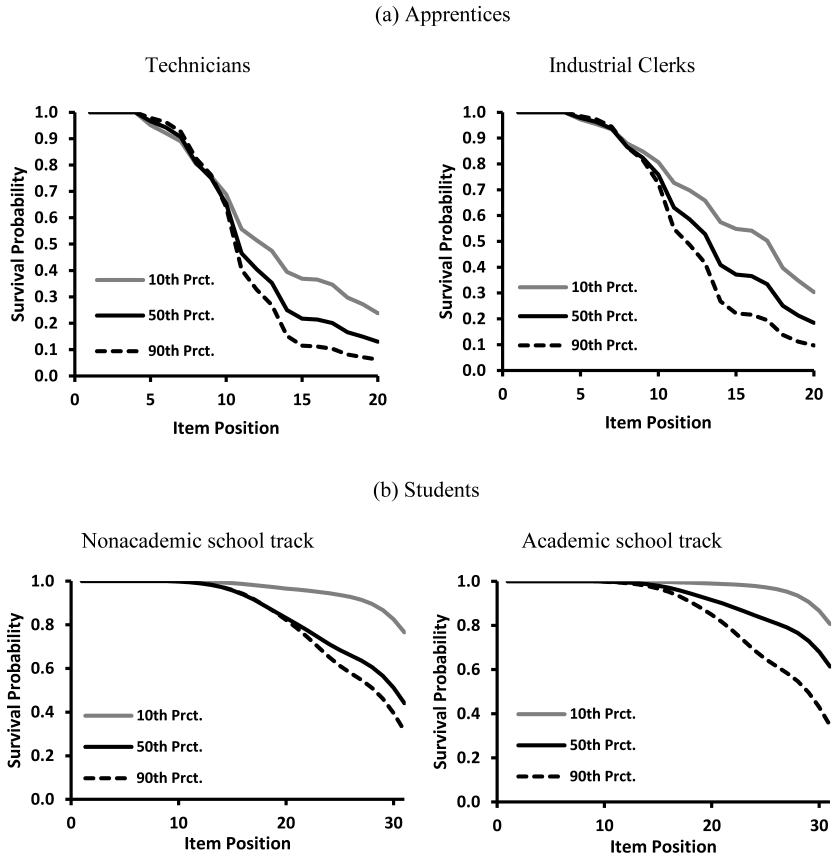

Figure 4.3: Survival functions determined for the 10th, 50th, and 90th percentiles (Prct.) of the joint proficiency distribution, determined for subgroups on the basis of the MDSEM.

and regions where NRIs did not depend on the variables considered. Finally, as we have demonstrated, the MDSEM allows for a simple test of the differential onset of NRIs as a function of the covariates, while simultaneously controlling for the proficiency variable.

Investigating the Onset of NRIs in a Sample of 7th-Grade Students

Sample and Procedure

The second sample consisted of seventh-grade students and was taken from the study "Learning Processes, Educational Careers, and Psychosocial Development in Adolescence and Young Adulthood" (BIJU; e.g., Köller, 1998). The test contained 31 items and was

administered with a time limit of 20 minutes. There were two booklets and the order of the first nine items was permutated. Both booklets were distributed randomly throughout the sample. Because of the item scrambling at the beginning of the test, 15 cases in which NRIs occurred before the 10th position were excluded prior to all analyses. Cases with less than three valid responses in the whole test were not included in the analyses. We considered data from N = 5,587 cases, of which 47% were in the academic track and 53% attended a nonacademic track school. On average, students were M = 12.71 years old (SD = 0.65).

In this application, we investigated the differential onset of NRIs with respect to school track because school track comparisons are at the core of many large-scale studies. Achievement differences between academic and nonacademic track students in Germany are well documented (e.g., Prenzel, Sälzer, Klieme & Köller, 2013). Some research indicates that academic track students take achievement tests more seriously (Baumert & Demmrich, 2001), which leads to the question of how track differences are related to the onset of NRIs. Köhler et al. (2015a) found that, in the ninth grade, students in the academic track had fewer NRIs than students in nonacademic tracks. However, the authors did not find substantive differences for students in the fifth grade. Hence, the question considered in this application was whether a differential onset of NRIs existed and, if so, whether this might influence the comparability of track differences in mathematics proficiency for students in the seventh grade.

Results

Descriptive Results. The baseline hazard function is given in Figure 4.1b; it showed a monotone increase in the proportion of students with NRIs. In contrast to the first application, the hazard probabilities were smaller, and there were no clear peaks in the function.

Nonparametric Representation of the Distribution of Steps Variable. The model fit statistics are presented in Table 4.1 (lower part). As shown there, the model with $K = 4$ classes achieved the best fit in terms of the BIC. The discrimination parameter of response indicators was estimated to be $\hat{\lambda} = 0.45$ ($SE = 0.020$), which means that the estimated probability of responding to items around the support points was less steep than in the sample of apprentices. This finding indicated that test takers could not be clearly classified by the points in the item sequence at which they were most likely to switch to NRIs.

The support points and posterior class proportions are summarized in Table 4.2 (lower part). The support points appeared to be concentrated at the end of the test. Two of the four support points were estimated to be higher than the maximum test length. However, since the support points were estimated at the inflection points of the probability of NRI onset across item positions, these two latent classes still had a probability greater than zero of showing NRIs with estimated survival probabilities (cf. Equation 4.10) of .50 for $c = 3$ and .82 for $c = 4$. These two latent classes had the highest class proportions, altogether, the majority of test takers had high probabilities of completing the test.

Comparison with the Model of Glas and Pimentel 2008. Similar to the sample of apprentices, we compared the MDSEM (without the school track covariate) with the model of Glas and Pimentel (2008). The model fit indices indicated that the MDSEM had a superior fit (MDSEM, 73 parameters: AIC = 185,245; BIC = 185,729; Glas & Pimentel, 66 parameters: AIC = 185,547; BIC = 185,984). The item parameters were nearly identical in both models (item discriminations: MAD = 0.008; item difficulties: MAD = 0.001). Also, the estimates of the variance of the proficiency variable were quite similar, with $\hat{\sigma}^2_\theta = 0.82$ ($SE = 0.022$) for the MDSEM and $\hat{\sigma}^2_\theta = 0.83$ ($SE = 0.022$) for the model of Glas and Pimentel (2008).

In both models, the relationship between proficiency and the onset of NRIs was negative, indicating that, higher levels of proficiency were related to earlier onsets of NRIs. In the model of Glas and Pimentel (2008), this effect was strong, with a standardized regression weight of $\hat{\gamma}^{\text{stnd}}_{\xi,\theta} = -.78$ ($SE = 0.179$, $p \leq .001$). This was also true for the MDSEM, although the effect was not linear; overall, higher levels of proficiency were associated with earlier onsets of NRIs. The multinomial logit coefficients of the regression of the latent classes on proficiency as well as the predicted class probabilities for low, medium, and high proficiency levels are displayed in Table 4.3 (lower part). The MDSEM clearly predicted that respondents with higher proficiency were more likely to show an earlier onset of NRIs, whereas the majority of low proficiency students were expected to belong to class $c = 4$.

Differential Onset of NRIs. In the last set of analyses, we evaluated whether school track (academic vs. nonacademic) was related to the onset of NRIs, while conditioning on proficiency. The estimated effect of school track (1 = *academic track*, 0 = *nonacademic track*), without conditioning on θ, on the onset of NRIs indicated that students in the academic track were more likely to show earlier onsets of NRIs [omnibus test of significance of coefficients: Wald's $\chi^2(df = 3) = 11.35$, $p < .05$]. However, this result could be a consequence of group differences in average proficiency levels, because students of higher proficiency were found to be more likely to show an earlier onset of NRIs. Therefore, in a next step, we evaluated the full MDSEM with proficiency and school track as predictors. Compared with a constrained MDSEM, where the direct effects of school track on latent classes were fixed to zero, we found a significant result for the LRT [$\chi^2(df = 3) = 61.3$ ($p \leq .001$)], indicating that school track was related to the onset of NRIs. Students in the academic track were found to have higher proficiency levels ($\hat{\gamma}_{\theta,x} = 1.12$, $SE = 0.023$, $p \leq .001$).

The multinomial logit coefficients are displayed in Table 4.4. When θ was included in the model, the effects of school track on latent classes were negative. For students with comparable levels of proficiency, those in the academic track were more likely to have later onsets of NRIs. Hence, the group differences in the onset of NRIs changed their direction once proficiency was controlled for.

The group-specific survival curves for different levels of proficiency (in the combined sample) are displayed in Figure 4.3b. Compared to the results derived on the basis of the previous sample, the survival curves were quite smooth. Within each group, test takers with lower proficiency tended to show later onsets of NRIs. However, in the academic track subsample, the survival curves were higher at each level of proficiency, but differences were clearly most accentuated for students with intermediate levels of proficiency.

Summary

In the second application of the MDSEM, we used an example recurring on a typical large-scale study. We investigated whether groups differed in the onset of NRIs. A differential onset of NRIs could indicate a possible violation to the assumption of equivalent proficiency scores. Our results provide an indication that students in the different school tracks showed a differential onset of NRIs. The analyses uncovered that students of average proficiency showed the most pronounced differences in NRIs, indicating a serious threat to the validity of group comparisons. However, in this case, we do not believe that the differential onsets of NRIs were of practical relevance for judging the size of group differences in proficiencies because these differences were so large that biases caused by differential NRIs would not change this picture substantially.

The results obtained in the second sample document the flexibility of the suggested approach to also capture the onset of NRIs in a less extreme situation. In this sample, the hazard function shows a rather smooth, nearly monotonous increase in the probability of

the onset of NRIs (see Figure 4.1b) instead of the peaked form in the sample of apprentices (see Figure 4.1a). Therefore, in this application, the estimated discrimination parameter λ was smaller than in the sample of apprentices, which leads to smoother survival functions. Nevertheless, even in such a situation, the MDSEM was capable of differentiating between the regions of a test in which the predictors do, or do not, have effects on the onset of NRIs.

Although both datasets differed with respect to age and educational background of the participants, as well as with respect to the scope of the assessments considered, the MDSEM showed a better fit than the model proposed by Glas and Pimentel (2008). Thus, for both datasets, it seemed plausible to relax the assumption of a linear development of the logit of hazards as well as the assumption of a constant effect of proficiency on the onset of NRIs.

Simulation Study

In this section we report the results of a simulation study that was conducted in order to study the behavior of the MDSEM in the presence of a high amount of NRIs. We examined the MDSEM's capability of uncovering (1) item parameters (i.e., discriminations, α, and difficulties, β, Equation 4.2), (2) structural parameters pertaining to the relationship between a covariate x and the proficiency variable θ (i.e., $\gamma_{\theta,x}$, Equation 4.14) and the variance of θ conditional on x (i.e., $\sigma^2_{\zeta_\theta}$, Equation 4.14), and (3) the survival function $P\left(S_i \geq m|\theta_i, x_i\right)$ (Equation 4.16). We simulated item responses and response indicators for a test with $J = 30$ items administered to two groups (variable x) of equal size ($N =$ 1,000 per group). Three conditions with 100 replications per condition were examined. In the first condition, NRIs depended solely on group membership. In the second condition, NRIs were only affected by the proficiency variable θ. Finally, in the third condition, NRIs

depended on both characteristics. Except when constrained to be zero, the effects of x and θ on the onset of NRIs were held constant across conditions. The datasets were generated in such a way that all subjects had complete data on the first four items, and approximately 33% of the item responses were missing in all conditions.

The probabilities of NRI onsets were generated via a nonnormally distributed steps variable, δ (Equation 4.8) with λ fixed to one in the data-generation process. Nonnormality in δ was generated by means of a mixture of $K = 6$ univariate normal distributions with proportions $\pi_1 = .12$, $\pi_2 = .10$, $\pi_3 = .21$, $\pi_4 = .28$, $\pi_5 = .16$, and $\pi_6 = .14$, means $\mu_1 = 7$, $\mu_2 = 12$, $\mu_3 = 17$, $\mu_4 = 22$, $\mu_5 = 27$, and $\mu_6 = 35$, and variances $\sigma_1^2 = 1.00$, $\sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 2.25$, and $\sigma_6^2 = 12.25$. The relationships of δ with θ and x were generated via a multinomial logit model (Equation 4.15) with parameters $\gamma_{1,\theta}$ = -2.00, $\gamma_{2,\theta} = 0.50$, $\gamma_{3,\theta} = 1.50$, $\gamma_{4,\theta} = 2.00$, and $\gamma_{5,\theta}$ = -0.25 for θ, and $\gamma_{1,x}$ = -2.00, $\gamma_{2,x}$ = -0.75, $\gamma_{3,x}$ = -0.75, $\gamma_{4,x}$ = -2.00, and $\gamma_{5,x} = 2.00$ for x. Intercepts were set in such a way that the proportions π_1 to π_6 were equal across conditions.

The discrimination parameters for item responses α were generated according to a uniform distribution ranging between 0.5 and 1.5 (mean = 1), whereas the difficulty parameters β followed a standard normal distribution (mean = 0). We centered x to its mean and fixed the intercept of the structural regression part (Equation 4.14) to zero. The effect of x on θ was set to 0.5, and $\sigma_{\zeta_\theta}^2$ was set to 0.937. These values imply that θ had a variance of one and a mean of zero, so that the effect of x can be directly interpreted as a standardized effect.

In each condition the data was analyzed via the MDSEM with $K = 6$ classes, and a 2PL model in which the response indicators were ignored. In both models, θ was regressed on x, and both models were identified by constraining the mean of the item discriminations to one. In light of the previous results, we expected the MDSEM to provide less biased parameters than a 2PL ignoring NRIs in conditions in which NRIs were nonignorable (i.e.,

conditions two and three). In condition one, we expected both models to perform equally well because the missing-data process was completely due to x. Because of its flexible nature, we expected the MDSEM to provide unbiased estimates of survival functions for various combinations of θ and x (10th, 50th, and 90th percentiles of the θ-distribution for each value of x).

Results. Table 4.5 provides the bias (i.e., the difference between average estimates and population values) of the parameters $\gamma_{\theta,x}$, and $\sigma^2_{\zeta_\theta}$, as well as the coverage rates of the parameter estimates (i.e., the proportion of parameter estimates whose 95% confidence intervals included the population values). The results for the 2PL model that ignored NRIs show that this model provided essentially unbiased structural parameters with good coverage rates only in the first condition. In this model, the variability of θ was underestimated in conditions two and three. Furthermore, the regression weight $\gamma_{\theta,x}$ was clearly biased in the third condition. The bias in the estimate of $\gamma_{\theta,x}$ was lower in the second condition, where the model also provided an acceptable coverage rate. In contrast, the MDSEM provided virtually unbiased structural parameters that were accompanied by good coverage rates in all conditions studied.

Figure 4.4 provides scatter plots of the population values and the average item parameter estimates. Both models provided almost identical estimates that were virtually unbiased in the first condition. In the second and third conditions, where NRIs depended on θ, the MDSEM provided more accurate estimates for the items' discrimination parameters. In addition, the estimates of item difficulties appeared to be somewhat more accurate in the MDSEM, although the parameters provided by the standard 2PL were not strongly biased.

The last issue approached was the recovery of survival functions. The MDSEM correctly identified the variables not related to the onset of NRIs. Type I error rates for the

Table 4.5: Population Values, Parameter Bias, and Coverage Rates for Structural Parameters Provided by the 2PL Ignoring NRIs, and the MDSEM Separated by Conditions (Condition 1: NRIs Affected by x; Condition 2: NRIs Affected by θ; Condition 3: NRIs Affected by x and θ).

		2PL		MDSEM	
	Population	Bias	Coverage	Bias	Coverage
Condition 1					
$\gamma_{\theta,x}$	0.500	0.001	1.00	0.001	1.00
$\sigma^2_{\zeta_\theta}$	0.937	0.011	.99	0.013	.99
Condition 2					
$\gamma_{\theta,x}$	0.500	-0.043	.98	0.005	1.00
$\sigma^2_{\zeta_\theta}$	0.937	-0.132	.18	0.028	.98
Condition 3					
$\gamma_{\theta,x}$	0.500	-0.093	.50	0.003	1.00
$\sigma^2_{\zeta_\theta}$	0.937	-0.120	.28	0.022	1.00

Notes. $\gamma_{\theta,x}$ = regression weight of x predicting θ, $\sigma^2_{\zeta_\theta}$ = variance of θ conditional on x.

multinomial logistic regression weights of θ were close to the nominal rate of .05 (range .06 to .02) in condition one, and the same pattern was found for the Type I errors of the logistic regression weights of x in condition two (range .08 to .03). The survival functions provided by the model for the 10th, 50th, and 90th percentiles of the θ-distribution for both values of x are presented in Figure 4.5. As can be seen, the survival functions were virtually unbiased.

Summary. The results clearly show the advantages of the MDSEM. The model accurately estimated the survival functions in each condition studied, thereby underscoring the MDSEM's utility for examining the determinants of test takers' onset points of NRIs. In addition, our results show that the MDSEM reduced biases in parameters caused by non-ignorable missing data. The MDSEM provided parameter estimates identical to the 2PL in a situation in which the missing-data process was accurately modeled by the inclusion of the covariate (i.e., condition one; Glas et al., 2015). In the conditions in which the onset of NRIs also depended on proficiency, the MDSEM provided unbiased estimates of

the structural parameters, whereas the standard 2PL did not. In these conditions, the MDSEM produced more accurate item parameters, although the bias was also relatively low in the case of the conventional 2PL model that ignored NRIs.

Discussion

In educational assessments, one concern is whether the amount of NRIs is related to the proficiency being measured. Such relationships are considered to be indicative of MNAR patterns, which means that not accounting for such relationships could induce bias in the estimates of item parameters and students' proficiencies (Ludlow & O'Leary, 1999; for a review, see: Pohl & Carstensen, 2013). However, an often overlooked point is the possible relationship between NRIs and the student characteristics that are at the core of comparative studies. As we have argued in this article, situations in which such relationships cannot be accounted for by the students' proficiencies indicate a differential onset of NRIs that can be regarded as a threat to the validity of group comparisons. However, whether the differential onset of NRIs is treated as an indication of a threat to the validity of group comparisons, or whether it is treated as a key outcome in its own right, depends on the goals of the study.

Following this line of reasoning, we have presented the MDSEM as a flexible semiparametric approach that can be used for examining the differential onset of NRIs. Our model stands in close relationship with the approach suggested by Glas and Pimentel (2008) but relaxes some of its implicit assumptions, including the parametric distribution of the steps variable that assesses the onset point of NRIs, and the proportional hazards assumptions used for assessing the relationships of NRIs with the proficiency variable. The MDSEM proved valuable for determining the regions in which the NRIs were related to the explana-

tory variables, whereas this is not possible in the model proposed by Glas and Pimentel (2008).

The MDSEM has some similarities with the GDM suggested by Köhler et al. (2015b) for modeling the possibly nonnormal distribution of proficiency and the tendency to omit item responses. In contrast to the GDM, in the MDSEM, only the distributional assumptions for the steps variable are relaxed, while the proficiency variable is still assumed to be normally distributed. Furthermore, in our model, the categorization of δ is based on freely estimated support points, whereas these are defined in advance in the case of the GDM. The MDSEM is conceptually different because it allows NRIs to be predicted by proficiency and other covariates, whereas Köhler et al.'s (2015b) GDM allows only the relationship between proficiency and the tendency for omissions to be examined. In addition, the MDSEM has similarities with the model of Bacci and Bartolucci (2015), which uses freely estimated support points for proficiencies as well for the latent tendency to omit items. However, our model relies on a smaller number of parameters because we specified a variable that can be clearly interpreted as a steps variable, thereby making the interpretation of the model in real applications easier.

In summary, the MDSEM is easy to implement with conventional software packages, and it provided a better description of the datasets considered in this article than the model of Glas and Pimentel (2008). The MDSEM facilitates a straightforward test of the differential onset of NRIs by means of the LRT, and enables the presentation of these effects in a manner that can be easily understood by using the survival function borrowed from discrete-time event history analysis (Allison, 2014). As such, we believe that the method will prove useful in real applications concerned with the phenomenon of the differential onset of NRIs.

Furthermore, as we have shown in the simulation study, the MDSEM proved valuable for optimizing parameter estimates in the presence of MNAR patterns that were caused

by NRIs. Compared to the 2PL model that ignored NRIs, the MDSEM clearly reduced biases in the variability of the proficiency variable and in group differences. As such, the MDSEM appears to be not only a valuable tool for examining whether NRIs are a threat to the validity of group comparisons, but also a model that helps to prevent such biases. However, this issue warrants further investigation. In particular, further studies should examine whether the MDSEM proves a viable alternative to existing models (e.g., Glas & Pimentel, 2008), as we think it does.

Future Developments

Although the MDSEM is highly flexible, it still includes assumptions, some of which can be easily relaxed. First, our hypothesis about the differential onset of NRIs was restricted to uniform effects, which means that the model assumed that respondents at all levels of proficiency were equally affected by this effect. Such a specification is common in other areas, for example, in studies investigating differential item functioning (DIF; Holland & Wainer, 1993). Following the DIF literature, the MDSEM could be extended to consider nonuniform effects by allowing the covariates to interact with the proficiency variable. Such models can be implemented in the case of categorical covariates by means of multigroup MDSEMs. Such an approach would make it possible to examine whether the effects of proficiency on the onset of NRIs differ between groups. In our opinion, evidence for an absence of the differential onset of NRIs would require group-invariant relationships between the proficiency variable and NRIs, as well as an absence of effects of the covariates on the onset of NRIs. Therefore, we decided to focus on uniform effects that can be interpreted more easily. However, extensions of the MDSEM that include interactions between the proficiency variable and covariates may be an interesting topic for further investigations.

A further restriction of the MDSEM is that it assumes the proficiency variable to be normally distributed and linearly related to the covariates. Given that these are standard

assumptions in continuous latent variable models, we do not consider them to be a general shortcoming of the MDSEM. However, similar to the GDM proposed by Köhler et al. (2015b), in the case of omitted items, the distributional assumptions regarding the proficiency variable could be relaxed. The merits of relaxing the MDSEM should be clearly examined.

In addition, the MDSEM assumes an invariance of the measurement model applied to the item responses across groups and across patterns of NRIs. The first restriction can be easily relaxed in the context of multigroup models. We decided not to pursue this point, mainly for pragmatic reasons and to enhance the ease of presentation. Relaxing the invariance assumption across patterns of NRIs, however, requires other types of models, such as the pattern mixture models suggested in the context of the missing-data literature (Little, 1993). In this context, our proposed semiparametric approach for assessing the steps variable could be used to stratify the sample according to the relevant patterns of NRIs (see Rose et al., 2010). In a next step, by drawing on the stratified sample, the invariance assumption could be relaxed. Further studies could consider this issue.

Conclusion

NRIs reflect a type of test-taking behavior that could be of interest in substantive research. As we have argued in this article, NRIs can either be regarded as a key outcome variable or can perhaps be treated as a threat to the validity of group comparisons of proficiency levels. In this article, we present the MDSEM as a flexible and easy-to-use approach for studying the onset of NRIs. As we have demonstrated, the MDSEM can be analyzed using standard software which might make this approach appealing for applied researchers.

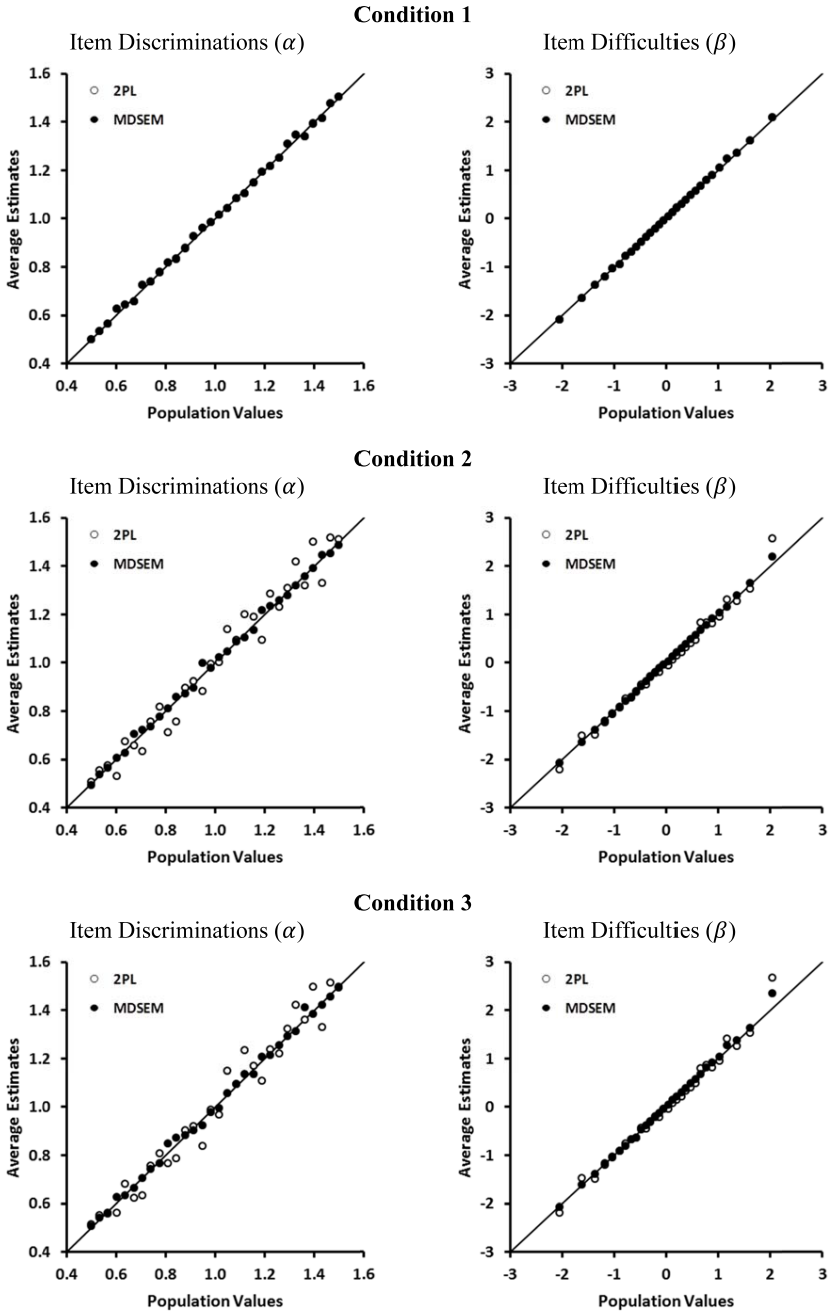

Figure 4.4: Estimated item parameters by corresponding population values for the 2PL ignoring NRIs and the MDSEM separated by conditions (Condition 1: NRIs affected by x; Condition 2: NRIs affected by θ; Condition 3: NRIs affected by x and θ).

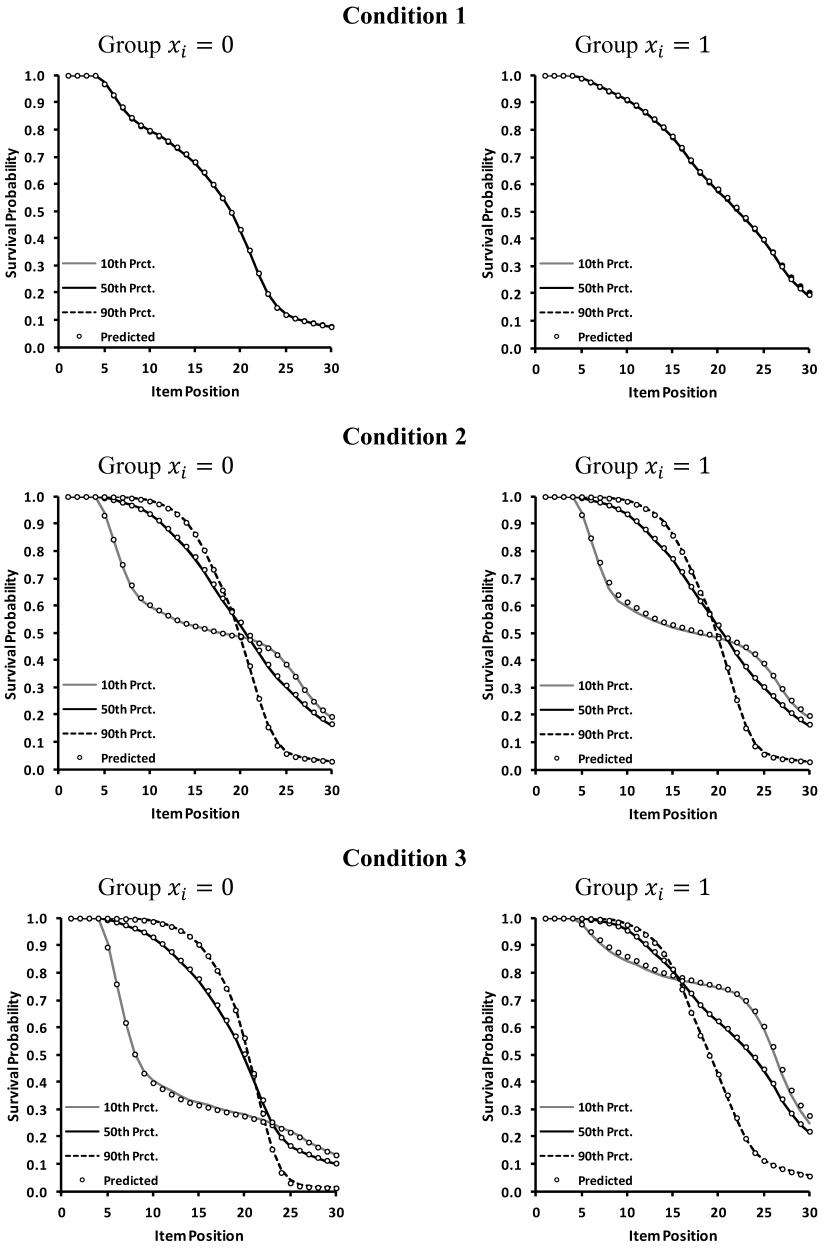

Figure 4.5: Population and estimated survival functions for the MDSEM separated by conditions (Condition 1: NRIs affected by x; Condition 2: NRIs affected by θ; Condition 3: NRIs affected by x and θ)

Kapitel 5

Studie 2:

Modellierung des Leistungsabfalls mittels Mischverteilungsmodellen

Die finale Version dieses Kapitels wurde in *Large-scale Assessments in Education* veröffentlicht, die Referenz lautet:

List, M. K., Robitzsch, A., Lüdtke, O., Köller, O. & Nagy, G. (2017). Performance decline in low-stakes educational assessments: Different mixture modeling approaches. *Large-scale Assessments in Education*, *5*(15), 1–25. Doi: 10.1186/s40536-017-0049-3

URL: http://rdcu.be/x0ia

Die vorliegende Fassung des Artikels zur zweiten Studie weicht aufgrund der editorischen Überarbeitung zur Drucklegung geringfügig von der publizierten Fassung ab.

Diese Arbeit nutzt Daten der Längsschnittstudien LAU und/oder KESS. Beide Datensätze wurden von der Freien und Hansestadt Hamburg durch die Behörde für Schule und Berufsbildung zwischen 1995 und 2012 generiert und dem Wissenschaftlichen Konsortium MILES (Methodological Issues in Longitudinal Educational Studies) für einen befristeten Zeitraum zur vertieften Bearbeitung wissenschaftlicher Fragestellungen zur Verfügung gestellt. MILES wird am Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) koordiniert.

Performance Decline in Low-Stakes Educational Assessments: Different Mixture Modeling Approaches

Abstract

In low-stakes educational assessments, test takers might show a performance decline (PD) on end-of-test items. PD is a concern in educational assessments, especially when groups of students are to be compared on the proficiency variable because item responses gathered in the groups could be differently affected by PD. In order to account for PD, mixture item response theory (IRT) models have been proposed in the literature. In this article, multigroup extensions of three existing mixture models that assess PD are compared. The models were applied to the mathematics test in a large-scale study targeting school track differences in proficiency. Despite the differences in the specification of PD, all three models showed rather similar item parameter estimates that were, however, different from the estimates given by a standard two parameter IRT model. In addition, all models indicated that the amount of PD differed between tracks, in that school track differences in proficiency were slightly reduced when PD was accounted for. Nevertheless, the models gave different estimates of the proportion of students showing PD, and differed somewhat from each other in the adjustment of proficiency scores for PD.

Keywords: *Educational Assessments, Mixture IRT Models, Performance Decline, Group Comparisons, Aberrant Response Behavior*

Introduction

One main purpose of large-scale assessments (LSAs) is to provide policy-makers and educational institutions with information about students' proficiency. Having no personal consequences, educational assessments are low-stakes tests for the test takers (Baumert & Demmrich, 2001; DeMars, 2000; Penk, Pöhlmann & Roppelt, 2014; S. L. Wise & DeMars, 2005) and, therefore, some test takers might not invest full effort throughout the test, resulting in an underestimation of their true proficiency levels. As a consequence, the estimates of proficiency scores and the inferences based on them (e.g., group differences) are likely to be biased (Eklöf, 2010; S. L. Wise & DeMars, 2005; S. L. Wise & Kong, 2005).

A common observation in educational LSAs is that the probability of a test taker giving a correct response decreases for items at the end of the test (Wu, 2010). On the test taker's side, this can be viewed as a performance decline (PD), which can be considered as a type of *aberrant response behavior* reflected in unexpectedly high rates of incorrect or omitted responses for end-of-test items (Cao & Stokes, 2008; Schnipke & Scrams, 1997; Suh et al., 2012). Whereas, in high-stakes tests, aberrant response behavior on end-of-test items is often attributed to test speededness (Bolt et al., 2002), in low-stakes tests, it is assumed that aberrant response behavior is related to a decline in test-taking effort (S. L. Wise & Kong, 2005). In order to explore PD in educational assessments, mixed-effects models (De Boeck & Wilson, 2004) have been proposed to analyze item position effects by comparing the probabilities of a correct response when the item is presented in different positions (Debeer & Janssen, 2013). An alternative strategy proposes using mixture models with categorical latent variables in order to identify the latent classes of test takers who differ in their test-taking behavior (Mislevy & Verhelst, 1990; Rost, 1990). Among others, Bolt, Cohen, and Wollack (2002), Yamamoto (1995), and Jin and Wang

(2014) introduced mixture models to separate test takers who show aberrant response behavior that corresponds to PD from test takers who respond to all items with full effort.

The aim of our study was to examine the utility of three mixture models to handle PD in low-stakes tests and to explore the differences and similarities in the conclusions drawn from these models. In this study, we applied the mixture models to investigate PD in the mathematics test of a German LSA. First, we explored the differences between the PD of test takers attending different school types. Next, we investigated whether PD affects the estimation of group differences in proficiency. Then, we compared existing mixture models with regard to the differences and similarities in their estimation of PD. Furthermore, we demonstrate how to fit these models using standard software such as Mplus (L. K. Muthén & Muthén, 1998-2012).

This article is organized as follows. First, we will provide an overview of the research on PD in educational assessments. We will then proceed to present three mixture models for PD that were extended to multigroup settings. After that, we will apply these models to a low-stakes mathematics test in order to compare their performance and parameter estimates. Finally, we will present and discuss the differences and similarities of these models regarding their measurement of PD and estimated group differences in proficiency.

Performance Decline in Educational Assessments

If PD is present, the probability of providing a correct response does not depend solely on item parameters and a person's proficiency. Simulation studies have shown that parameter estimates of end-of-test items are biased if PD is not taken into account (Oshima, 1994; Suh et al., 2012). This is of particular concern when parameter estimates of items are going to be treated as known in other applications, for example, in adaptive testing, as part of a calibrated item pool, or for test construction purposes (Davey & Lee, 2011; van Barneveld, 2007). In addition, test scores obtained under PD conditions are likely to be lower than

they would be if the test taker had invested full effort throughout the test. Thus, when PD is ignored, proficiency scores are underestimated – the stronger the PD effects are and the more test takers experience PD, the stronger the underestimation of the sample's average proficiency will be.

Several authors have studied associations between PD and test takers' characteristics, such as ethnicity, language skills, or gender. Bolt et al. (2002) analyzed a high-stakes college placement test for mathematics and found that the number of students showing PD differed between different ethnic groups. Yamamoto and Everson (1995) assessed PD in a high-stakes reading comprehension test for university students. They also found differences between the PD of different ethnic groups. Furthermore, they found that nonprimary-language speakers showed an earlier onset of PD than primary-language speakers. For a high-stakes reasoning test, Schnipke and Scrams (1997) also found that nonprimary-language speakers were more strongly affected by PD than primary-language speakers.

For low-stakes tests, differences in PD have been reported on a country level for the PISA assessments (Debeer et al., 2014; Hartig & Buchholz, 2012; Jin & Wang, 2014). G. Nagy, Lüdtke, and Köller (2016) also found PD effects to differ between school types in the German PISA 2012 study. Furthermore, male test takers have been found to show more guessing behavior and lower levels of test-taking motivation than female test takers (DeMars, Bashkov & Socha, 2013).

Due to the differential effects found, these results imply that ignoring PD might affect the estimations of average proficiency levels in a group-specific way. Thus, when investigating group differences in proficiency, the extent to which the differences found might be caused by differences in PD instead of true differences in proficiency should be explored (DeMars et al., 2013; Denis & Gilbert, 2012; Mittelhaëuser et al., 2015a).

Mixture Models of Performance Decline

Mixture *item response theory* (IRT; e.g., Embretson & Reise, 2000) models can be used to identify test takers showing PD. We will refer to these models as *mixture PD models* in the rest of this article. In general, mixture models assume that the population consists of subgroups, for which the model parameters differ in particular ways (Mislevy & Verhelst, 1990; Rost, 1990). These subgroups are also referred to as *latent classes* since they are not observed. Based on their individual responses, the probabilities for each test taker of belonging to one of these latent classes are estimated along with the other model parameters. Mixture PD models consist of two or more latent classes, where one class represents test-taking behavior that reflects full effort throughout the test (the *no-decline class*), and the other classes represent test-taking behavior that reflects PD (the *decline classes).*

Of these, two models, the two-class mixture model of Bolt et al. (2002) and the HYBRID model of Yamamoto (1995) have been applied to several empirical and simulated data sets (e.g., Boughton & Yamamoto, 2007; Cao & Stokes, 2008; Hailey et al., 2012; Mittelhaëuser et al., 2013; Mittelhaëuser et al., 2015a; Suh et al., 2012; Wollack et al., 2003; Yamamoto & Everson, 1997). Bolt et al. (2002) developed their two-class mixture model to reduce the bias in item parameters that is caused by test speededness in high-stakes tests. The aim of the model is to identify the group of test takers who run out of time and show PD (i.e., the decline class). PD is reflected in higher difficulty parameters for end-of-test items. The item parameters of the no-decline class are expected to be unbiased. Later on, the model was also used to model differences in test-taking motivation in low-stakes tests in order to separate motivated from unmotivated test-taking behavior (e.g., Mittelhaëuser et al., 2015a).

The HYBRID model of Yamamoto (1995) was initially developed to model changes in test-taking behavior that occur under conditions of test speededness in high-stakes tests.

The HYBRID model focuses on determining the position in which a test taker switches from effortful response behavior to aberrant response behavior when running out of time. The item responses that reflect aberrant response behavior are ignored for the estimation of proficiency and item parameters; therefore, item parameter and proficiency estimates are expected to be unbiased. The HYBRID model has also been used to investigate PD in low-stakes tests (cf. Cao & Stokes, 2008).

Recently, Jin and Wang (2014) proposed a multiclass mixture model to account for PD. Similar to Yamamoto's (1995) HYBRID model, it makes it possible to determine the point where test takers alter their test-taking behavior. In contrast to the HYBRID model, Jin and Wang (2014) do not assume that test takers switch to entirely aberrant response behavior that is unrelated to proficiency. Rather, the authors model PD as a decline in test performance so that the probability of providing a correct response decreases after the onset of PD.

In a recent study, the HYBRID model (Yamamoto, 1995) and the two-class mixture model of Bolt et al. (2002) were compared with regard to their capability to reduce the bias in item parameters that is caused by test speededness (Suh et al., 2012). Both models were found to be equally capable of estimating the true item parameters and outperformed a standard IRT model that did not account for test speededness. In addition, in both cases, the results were not affected by the coding of the items not reached by the examinees (i.e., missing vs. incorrect). However, the focus of the study of Suh et al. (2012) was bias in item parameters; the effects of PD on the estimation of proficiency therefore need further investigation. Furthermore, the model of Jin and Wang (2014) has not yet been investigated thoroughly, and its performance has not yet been compared with that of the model of Yamamoto (1995) or the model of Bolt et al. (2002).

A General Mixture IRT Model of Performance Decline

In the three aforementioned mixture PD models, several common assumptions are made. It is assumed that test takers respond to the items in the order in which they are presented. It is further assumed that PD starts at some point in the test, that is, the test takers will have responded to at least the first item with full effort before their test-taking behavior reflects PD. Thus, a *switching point* (von Davier & Yamamoto, 2007) exists at which decline onset can be observed, dividing the item response vector of a test taker into two parts: in the first part, item responses depend solely on the test taker's proficiency and on the item characteristics. In the second part, item responses depend on a latent person variable underlying the aberrant response behavior that reflects PD. Depending on the definition of PD as specified in the particular model, aberrant response behavior may or may not be related to proficiency. For both parts, the relationship between item responses and the latent person variable can be modeled within the IRT framework. We assume that, before the switching point, a two-parameter logistic model (2PLM; Birnbaum, 1968) holds. However, after the switching point, we assume that item and person parameters might be different. Furthermore, we assume that some test takers show full effort throughout the test. Thus, the sample consists of a mixture of test takers with and without PD.

Based on these assumptions, a *general mixture PD model* can be described, which contains the aforementioned models as special cases. In the following, we describe such a model for the multigroup case, although the framework can be easily adapted for single group situations. According to the 2PLM, the probability of person p providing a correct response to item i depends on the person's proficiency (θ_p) and the item's slope and intercept parameter (α_i, β_i). The person's individual switching point (δ_p) equals the last item position before decline onset. Thus, the switching point can take values $\delta_p = 1, \ldots, I$, where I is the total number of items in the test. If $\delta_p = i$, PD will start after item i. If $\delta_p = 1$, PD will start after the first item. If $\delta_p = I$, PD will not occur at all in the test.

Let η_{pi} be the logit of the probability of correct response of person p to item i: $\eta_{pi} = \text{logit } P(X_{pi} = 1|\theta_p, \delta_p)$. For the general mixture PD model, the conditional item response probabilities are defined as

$$\eta_{pi} = \begin{cases} \alpha_i \cdot \theta_p + \beta_i, \text{ if } \delta_p \geq i, \\ \tilde{\alpha}_{ig} \cdot \tilde{\theta}_p + \tilde{\beta}_{ig}, \text{ if } \delta_p < i. \end{cases} \tag{5.1}$$

where θ_p denotes the proficiency variable, α_i is the slope parameter or item discrimination and β_i is the intercept parameter[19], both before the switching point, that is, for $\delta_p \leq i$. Note that the item parameters applied to responses before the switching point contain no group index g, so that we assume them to be invariant across groups, although this assumption can be relaxed.

The parameters $\tilde{\alpha}_{\text{ig}}$ and $\tilde{\beta}_{\text{ig}}$ denote the slope and intercept parameter in group g after the switching point, that is, for $\delta_p > i$, respectively, and $\tilde{\theta}_p$ denotes the person variable underlying aberrant response behavior. When a test taker shows PD, the proficiency variable θ_p can change to $\tilde{\theta}_p$, which would no longer be interpreted as proficiency. Equation 5.1 is a mixture 2PLM where the latent classes capture the magnitude of the individual PD. If $\delta_p = I$ for all test takers in the sample, Equation 5.1 is reduced to $\eta_{pi} = \alpha_i \cdot \theta_p + \beta_i$, which is the standard 2PLM.

The joint distribution of θ and δ is defined based on the assumption of how proficiency and switching point are related. Typically, θ is considered to be a normally distributed continuous variable. By definition, δ is considered to be a discrete variable. One way of specifying the joint distribution $P(\theta, \delta|G = g)$ is to consider the conditional distribution of θ with respect to δ in group g (i.e., $P(\delta|G = g)$), that is, $P(\theta, \delta|G = g) = P(\theta|\delta, G = g) \cdot P(\delta|G = g)$. Within each class and each group, a normal distribution of θ is assumed,

[19]Note that the item difficulty b_i is the negative of the ratio of intercept to slope parameter (Hambleton, Swaminathan & Rogers, 1991), that is, $b_i = -\beta_i/\alpha_i$.

that is, $\mathrm{P}(\theta|\delta = i, G = g) = N(\mu_{\mathrm{ig}}, \sigma^2_{\mathrm{ig}})$. Typically, in the case of many classes, restrictions are imposed on the means μ_{ig} and standard deviations σ_{ig} (e.g., Yamamoto & Everson, 1997).

The dependencies between θ and δ as assessed by multigroup mixture PD models could be of substantive interest in real applications. For example, research suggests that students with lower proficiency estimates are more likely to show aberrant response behavior, as reflected in PD (Bolt et al., 2002; De Boeck et al., 2011; S. L. Wise & Kong, 2005); mixture PD models make it possible to examine such hypotheses. In addition, researchers might expect that the strengths of the relationship between θ and δ differ between groups (e.g., groups assessed in low-stakes vs. high-stakes conditions). Again, the multigroup setup of mixture PD models allows such hypotheses to be explicitly tested.

In addition, the multigroup mixture PD models provide estimates for the proportions of decline classes in group g, $\pi_{\mathrm{ig}} = P(\delta = i|G = g)$ for $i < I$, including the proportions of test takers not affected by PD, $\pi_{\mathrm{Ig}} = P(\delta = I|G = g)$, thereby allowing researchers to examine whether the proportion of test takers not affected by PD differs between groups, and to investigate group differences in the prevalence of earlier and later onsets of PD.

However, the assessment of latent class probabilities, π_{ig}, might not succeed in mixture PD models that contain many latent classes without imposing additional constraints. In order to solve this problem, Cao and Stokes (2008) proposed a cumulative probability function of switching points, which was also employed by Jin and Wang (2014) in their mixture PD model. In the multigroup case, the function recurs on the probability of the no-decline class in group g, π_{Ig}, and the probabilities of the decline classes,π_{ig}, for $i < I$. Cao and Stokes (2008) assume that the cumulative probability function of δ_p depends on a shape parameter $\omega > 0$, so that

$$P\left(\delta_p \leq i|G = g\right) = \frac{i^{\omega_g}}{(I-1)^{\omega_g}}, \text{for } i < I \tag{5.2}$$

The shape parameter ω_g defines the form of the cumulative probability curve in group g: if $\omega_g = 1$, the probability increases linearly, if $\omega_g > 1$, the increase is convex, and if $\omega_g < 1$, the increase is concave. With Equation 5.2, the curve of π_{ig} is described as

$$\pi_{\text{ig}} = \frac{i^{\omega_g} - (i-1)^{\omega_g}}{(I-1)^{\omega_g}} \cdot (1 - \pi_{\text{Ig}}), \text{for } i < I. \tag{5.3}$$

In Equation 5.3, only two parameters are estimated per group, the proportion of the no-decline class π_{Ig} and the shape parameter ω_g; this reduces the number of parameters for long tests with many switching points. Because of its parsimony, we used Equation 5.3 in our study to model the probability distributions for the decline classes in mixture PD models with many classes. However, other functions could be used as well.

The cumulative probability function of δ_p (Equation 5.2) used in our application allows researchers to explicitly test hypotheses about group differences in the proportion of examinees affected by PD via the estimates of π_{Ig} (Equation 5.3). In addition, the function allows for a straightforward comparison of the ω_g-parameters that indicate whether the onset of PD occurs later (higher values of ω) or earlier in the test (lower values of ω). However, as the group-specific proportions of PD onset points π_{ig} is a rather complex function of ω_g and π_{Ig} (Equation 5.3), we suggest using graphical aids for comparing the distribution of onset points across groups.

Finally, it should be noted that the multigroup mixture PD models make it possible to assess not only the group-specific distributions of the onset points of PD but also the magnitude of PD in each group. However, the magnitude of PD depends on the specific restrictions imposed on the general mixture PD model (Equation 5.1): in some models, the magnitude of PD is expressed via the $\tilde{\beta}_{\text{ig}}$-parameters (Bolt et al., 2002; Yamamoto, 1995), whereas, in others, the magnitude of PD is quantified by the person variable $\tilde{\theta}_{\text{p}}$ (Jin & Wang, 2014).

Multigroup Mixture IRT Models of Performance Decline

By placing specific parameter restrictions, the general mixture PD model can be transformed into multigroup versions of each of the three aforementioned mixture PD models (Bolt et al., 2002; Jin & Wang, 2014; Yamamoto, 1995). Since PD is modeled differently in the specific mixture PD models, group comparisons for each mixture PD model involve different parameters. In the following, we will describe how the general mixture PD model can be extended to realize the mixture PD models of Bolt et al. (2002), Yamamoto (1995), and Jin and Wang (2014), allowing for multigroup comparisons.

The two-class mixture model of Bolt et al. (2002)

Bolt et al. (2002) proposed a mixture PD model with two latent classes (which will further be referred to as the 2PDM): while test takers in the no-decline class do not show PD, test takers in the decline class show lower performance on end-of-test items. Therefore, in the decline class, items appear to be more difficult than in the no-decline class, resulting in lower item intercept parameter estimates after the switching point. In their original formulation of the 2PDM, Bolt et al. (2002) suggested specifying the switching point i_0 in advance (see also De Boeck et al., 2011; Wollack et al., 2003), so that it refers to an item position up to which responses are expected to not be affected by PD. When put into the multigroup context, this specification implies that, in each group, the switching point is a dichotomous variable that can take two values, that is, $\delta_p = i_0$ for all test takers showing PD, and $\delta_p = I$ for all test takers not showing PD. Accordingly, in each group, there are two latent classes with proportions $\pi_{i_0 g} = P(\delta = i_0 | G = g)$ for the decline class and $\pi_{Ig} = P(\delta = I | G = g)$ for the no-decline class.

The model was first proposed as a mixture Rasch model with equal item discrimination parameters for all items. Recent applications have also considered extensions based on a 2PLM (Cao & Stokes, 2008; Suh et al., 2012), which we also used in our study. For

each group, we restricted the item discriminations to be equal in both latent classes (i.e., $\tilde{\alpha}_{\text{ig}} = \alpha_i$; see Cao & Stokes, 2008). After the switching point, item intercept parameters were allowed to change but item responses to still depend solely on the item parameters and proficiency. Thus, the person variable underlying the aberrant response behavior would still be regarded as proficiency, hence $\tilde{\theta}_p = \theta_p$. Within each group, item intercept parameters were constrained to be lower in the decline class (i.e., $\tilde{\beta}_{\text{ig}} \leq \beta_i$ for each g). Equation 5.1 can be altered to realize the multigroup version of the 2PDM with

$$\eta_{pi} = \begin{cases} \alpha_i \cdot \theta_p + \beta_i, \text{ if } \delta_p \geq i_0, \\ \alpha_i \cdot \theta_p + \tilde{\beta}_{ig}, \text{ if } \delta_p < i_0. \end{cases} \tag{5.4}$$

The mean proficiency is allowed to vary across groups and classes, while the standard deviation is held equal across latent classes, that is, $\theta \sim N(\mu_{\text{ig}}, \sigma_g^2)$. We assumed that groups could differ in the intercept parameters after the switching point, $\tilde{\beta}_{\text{ig}}$ (see Equation 5.4). In addition, we did not impose any constraints on the latent class proportions, $\pi_{i_0 g}$ and π_{Ig}.

Note that, in the common specification of the 2PDM, the (arbitrarily) specified switching point i_0 does not identify the item position in which PD first occurs. Rather, researchers are required to compare the estimates of β_i and $\tilde{\beta}_{\text{ig}}$ and to identify the position from which these estimates show meaningful deviations from each other. An alternative is to identify the point at which PD first occurs by comparing the model-data-fit of the 2PDM, specified with different values of i_0, and selecting the switching point that provides the best fit to the data at hand. This procedure has the advantage that many β_i-parameters that would otherwise be specified to be class- and group-specific (i.e., $\tilde{\beta}_{\text{ig}}$-parameters) are estimated on the basis of a larger number of item responses. The procedure can be extended to identify group-specific switching points, but the merits of such an approach seem to be limited. When a common switching point is assumed, researchers can inspect the results for group

differences in the onset of PD by comparing the estimates of $\tilde{\beta}_{ig}$ with the corresponding estimates of β_i. Results showing negligible discrepancies between $\tilde{\beta}_{ig}$- and β_i-parameters after the switching point i_0 indicate a later onset point of PD in this group.

Hence, the multigroup 2PDM makes it possible to (1) assess the relationship between proficiency and PD on the basis of the latent class means μ_{ig}, (2) estimate the proportion of examinees not affected by PD in each group (π_{Ig}), and (3) quantify the strengths of PD in each group by examining the discrepancies between $\tilde{\beta}_{ig}$- and β_i-parameters. In addition, the latter parameters also make it possible to (4) inspect the results for group differences in the onset points of PD.

The HYBRID model of Yamamoto (1995)

In his HYBRID model (further referred to as the HYBRID), Yamamoto (1995) assumes that, after the switching point, item responses no longer reflect test takers' proficiency. Instead, the probability of a test taker providing a correct response to items after PD onset is independent of proficiency; rather, it corresponds to an item-specific response threshold. While the model was originally proposed to model random guessing in high-stakes tests under speededness conditions, it can also be applied to other types of aberrant response behavior reflecting PD (e.g., Suh et al., 2012).

PD onset can occur in all item positions except the first one, hence, there are multiple decline classes, one for each item position, so that $\delta_p = 1, \ldots, I$ in each group. Within each group, the response thresholds can be specified to be either item-specific or equal for all items. The latter assumption is reasonable in cases where all items share a similar response format, such as multiple-choice items with the same number of options.

Since item responses after the switching point do not depend on proficiency, the slope parameter after the switching point is set to zero in each group ($\tilde{\alpha}_{ig} = 0$, for each g). Since the slope parameters after the switching point are set to zero, the item responses do not

depend on the person variable, $\tilde{\theta}_p$, meaning that the person variable underlying aberrant response behavior is not defined. To keep the model identified, we set $\tilde{\theta}_p = \theta_p$. Note that $\tilde{\theta}_p$ can be fixed to any other value, as it does not impact on η_{pi} after the switching point. The intercept parameter after the switching point is constrained to a common response threshold within each group, which we restricted to be the same for all items in all decline classes, that is, $\tilde{\beta}_{\text{ig}} = \tilde{\beta}_g$ for all i.

Thus, within each group, the probability of a correct response after PD onset is the same for all test takers and items regardless of proficiency (or any other person variable underlying aberrant response behavior) and regardless of the location of PD onset. Based on these specifications, the multigroup HYBRID can be derived by altering Equation 5.1 to

$$\eta_{pi} = \begin{cases} \alpha_i \cdot \theta_p + \beta_i, \text{ if } \delta_p \geq i, \\ 0 \cdot \theta_p + \tilde{\beta}_g, \text{ if } \delta_p < i. \end{cases} \tag{5.5}$$

The distribution of decline class probabilities within each group was assumed to follow the function described in Equations 2 and 3. We assumed that the mean of θ in each PD class assessed in each group is a function of δ, while the standard deviation of θ is the same for all latent classes but differs between groups. More specifically, we modeled the mean of the conditional θ distributions across latent classes as a linear function (Yamamoto & Everson, 1997) with the mean of the no-decline class, μ_{Ig}, as the intercept and a group-specific slope parameter ρ_g:

$$\mu_{\text{ig}} = \mu_{\text{Ig}} + \rho_g \cdot (I - i)\,. \tag{5.6}$$

The value of ρ_g in Equation 5.5 shows how θ and δ are related: if ρ_g is negative, average proficiency is lower, the earlier the PD onset. If ρ_g is positive, average proficiency would be lower for later PD onsets.

Taken together, the multigroup HYBRID makes it possible to assess the group-specific (1) relationships between proficiency and PD via the parameters ρ_g, (2) proportions of examinees not affected by PD by means of the latent class proportions π_{Ig}, and (3) strengths of PD effects that are governed by the response thresholds $\tilde{\beta}_g$. Compared to the multigroup 2PDM, the HYBRID model presented here allows for (4) a more finely-grained assessment of the group-specific onset points of PD by inspecting the group-specific cumulative probability functions of δ_p (Equation 5.2) and their parameters ω_g.

The multiclass mixture performance decline model of Jin and Wang (2014)

Jin and Wang (2014) proposed another multiclass mixture PD model (further referred to as the MPDM). Similar to the HYBRID, PD onset can occur after any item position throughout the test. In the MPDM, it is assumed that, after the switching point, θ_p is reduced according to a decrement function which depends on δ_p. Hence, when specified in a multigroup context, in the MPDM, PD is modeled via a group-specific change in the θ-variable instead of via changes in the intercept parameters as is the case in the 2PDM and the HYBRID model. More specifically, in the multigroup MPDM, it is assumed that, within each group, the value of the decrement function is the same for all test takers who have the same switching point. It is further assumed that, in all classes, the decrement is smaller when PD onset occurs later in the test.

Jin and Wang (2014) assume that, after PD onset, the item parameters are the same as before the switching point ($\tilde{\alpha}_i = \alpha_i$and $\tilde{\beta}_i = \beta_i$), but that the person variable underlying

the aberrant response behavior, $\tilde{\theta}_p$, corresponds to the difference between proficiency and the decrement for the respective switching point, that is, $\tilde{\theta}_p = \theta_p - \kappa_g \cdot (I - \delta_p)$. The multigroup MPDM can be formalized by altering Equation 5.1 to

$$\eta_{pi} = \begin{cases} \alpha_i \cdot \theta_p + \beta_i, \text{ if } \delta_p \geq i, \\ \alpha_i \cdot [\theta_p - \kappa_g \cdot (I - \delta_p)] + \beta_i, \text{ if } \delta_p < i. \end{cases} \tag{5.7}$$

As shown in Equation 5.7, the decrement in θ occurring after PD onset is a linear decreasing function across switching points with a group-specific decrement parameter $\kappa_g > 0$ as the slope so that, for the last decline class, the decrement is κ_g.[20] Note that Jin and Wang (2014) modeled θ and δ to be independent of one another. However, we suggest applying Equation 5.6 in the same manner as described for the HYBRID for two reasons. First, in this way, it is possible to empirically investigate whether the independence assumption that Jin and Wang (2014) propose holds. Second, we wanted to base our comparisons of the three mixture PD models on similar assumptions regarding proficiency and PD throughout all models. In our multigroup version of the MPDM, the proportions of decline classes were defined in the same way as for the HYBRID, with group-specific shape parameters ω_g (Equations 2 and 3).

Hence, it can be summarized that the multigroup MPDM shares some similarities with the HYBRID: It makes it possible to examine the group-specific (1) relationships between proficiency and PD (ρ_g-parameters), (2) proportions of examinees showing no PD (proportions π_{Ig}), and (3) distributions of PD onsets δ_p (Equation 5.2). However, the MPDM differs from the HYBRID because it assumes (4) that the magnitude of PD depends on δ_p, although the size of PD could be group-specific, and (5) that the responses affected by PD still depend on proficiency.

[20] Although Jin and Wang (2014) also discuss quadratic functions for the decrement function, in their empirical application, they found that a linear function was sufficient.

The Present Investigation

The aim of the present study was to examine the extent to which the multigroup mixture PD models make it possible to assess PD, and whether the conclusions about the presence and group differences in PD that can be drawn from these models differ. In addition, we examined whether accounting for group differences in PD affected the results of the group comparisons of students' proficiencies, and whether the mixture models envisaged differed in their estimates of group differences in proficiencies. Finally, we investigated whether the mixture PD models provided item parameters estimates that differed from those given by the multigroup 2PLM (Suh et al., 2012). To this end, we drew on a LSA of German students who worked on a mathematics test. We fitted the multigroup versions of the 2PDM, the HYBRID, the MPDM, and a standard multigroup 2PLM. We chose school track as a grouping variable, since school track comparisons are at the core of many large-scale educational programs, and school track has been found to be strongly related to PD, as reflected in item position effects (G. Nagy, Lüdtke & Köller, 2016).

In Germany, after attending primary school, children are assigned to different school tracks based on their school achievement. There is one academic school track (higher secondary school, *Gymnasium*) and there are several non-academic school tracks, including comprehensive (*Gesamtschule*), intermediate (*Realschule*), and lower secondary schools (*Hauptschule*), though the number of non-academic school tracks can vary (Pietsch & Stubbe, 2007). Achievement differences between students at academic and non-academic school tracks are known to be large (e.g., Prenzel et al., 2013). Furthermore, students at the non-academic tracks appear to take participation in LSAs less seriously (Baumert & Demmrich, 2001), possibly giving rise to stronger PD. As such, group differences in PD are likely, and group comparisons of students' proficiencies could be affected by group differences in PD (Mittelhaëuser et al., 2015a; G. Nagy, Lüdtke & Köller, 2016).

Research Questions

The analyses reported in this article examined both substantive and methodological research questions. From a substantive point of view, we applied the multigroup mixture PD models in order to examine (Q1) the existence of PD in item responses and track differences therein. Regarding the differences between tracks, we examined (Q2) the proportions of students not affected by PD, (Q3) the distributions of onset points of PD, (Q4) the magnitude of PD effects in each group, and (Q5) the relationships between PD behavior and students' proficiencies.

We expected to obtain the following results: regarding Q1, since the assessment considered was a low-stakes test, we assumed that our data would show some degree of PD and, thus, we expected all mixture PD models to provide a better fit to the data than the 2PLM. Regarding Q2 to Q4, in line with research on school track differences regarding item position effects (G. Nagy, Lüdtke & Köller, 2016) and test-taking motivation (Baumert & Demmrich, 2001), we expected the amount of PD to be larger for students attending the non-academic tracks. More specifically, we expected the proportion of students not affected by PD to be higher in the academic track (Q2). We did not feel able to derive detailed expectations about group differences in the onset point of PD (Q3), or in the magnitude of PD effects (Q4) because, to the best to our knowledge, such issues have not been investigated previously. We therefore treated Q3 and Q4 as open research questions to be examined in our application.

So far, only few studies have dealt with the relationship between PD and proficiency (Q5). However, as research on response times indicates that low test-taking effort is associated with low proficiency (S. L. Wise & DeMars, 2005; S. L. Wise et al., 2009), it seemed reasonable to assume that less proficient students would be more likely to show PD. We therefore expected that, within each school track, average proficiency would be lower, the earlier PD occurs.

Our second set of research questions targeted methodological issues. Here, we were interested in (Q6) whether accounting for PD affects the estimates of item parameters, (Q7) whether the multigroup mixture PD models differ in the conclusions that can be drawn from them about the prevalence of PD, and (Q8) the impact that PD has on the results of group comparisons of proficiency.

When PD occurs, the intercept parameters of end-of-test items are likely to be underestimated, whereas their slope parameters are likely to be overestimated when a standard 2PLM is employed (Bolt et al., 2002; Oshima, 1994; Suh et al., 2012). Thus, regarding Q6, we expected the estimates of the intercept parameters of end-of-test items in the no-decline class to be higher for the mixture PD models than for the 2PLM, whereas we expected to obtain the opposite result for the item discriminations. We expected this result to hold for all mixture PD models, as Suh et al. (2012) found that most mixture PD models showed quite similar behavior in this respect. Regarding Q7, we did not feel able to derive detailed expectations because the results provided by the models depend on their representation of PD. Therefore, we treated the question of whether the models presented here converge to similar conclusions as an open research question.

Regarding Q8, we expected that not accounting for PD would lead to an underestimation of the proficiency for the decline classes. Thus, we expected to obtain higher estimates of average proficiency when mixture PD models were employed (Q5a). Since PD was expected to be higher at the non-academic school tracks, we further expected the underestimation of proficiency to be more pronounced in the non-academic than in the academic group. As a consequence, accounting for PD was expected to result in smaller group differences in proficiency between both school tracks. However, the question of whether all models lead to a similar reduction in group differences in proficiency remained open.

Method

Sample and Test Design. We considered the mathematics achievement test of the first measurement point of a German large-scale longitudinal educational study, "Aspects of learning background and learning development", which was conducted in the federal state of Hamburg (Lehmann & Peek, 2011). Participation in the assessment was mandatory. The test had no individual consequences for the test takers and, hence, it can be regarded as a low-stakes test. The sample consisted of N = 12,182 students in the fifth grade at different school tracks, an academic track (n = 5,333 students) and two non-academic school tracks.

The mathematics assessment was presented as a paper-and-pencil test with a fixed item order, consisting of 30 multiple-choice items, scored as correct or incorrect. Missing item responses caused by omitted and not-reached items were also coded as incorrect. A distinction between incorrect and missing responses was not possible because the original item responses were not made available by the primary investigator. Although this means that PD effects were also influenced by the number of not-reached items, this issue does not appear to be of importance in mixture PD models. Suh et al. (2012) found the 2PDM and HYBRID model to perform equally well regardless of whether not-reached items were coded as incorrect or missing. Furthermore, Jin and Wang (2014) explicitly recommended coding not-reached items as incorrect in their MPDM.

Statistical Analyses. We applied the three mixture models (2PDM, HYBRID, MPDM) in the multigroup extensions presented and a multigroup 2PLM to the mathematics achievement test data. The grouping variable was school track and we considered two groups – students attending the academic school track, and students attending a non-academic school track. The models were compared by means of the Bayesian information criterion (BIC; Schwarz, 1978) and Akaike's information criterion (AIC; Akaike, 1987). Group differences

in PD were investigated by comparing the group-specific parameters for each model, and Wald tests were used to test for significant parameter differences between groups. Note that some parameters could not be compared across models; therefore, we investigated whether the patterns found in each model led to similar conclusions regarding PD.

Model Specification and Parameter Estimation. For model identification purposes, the mean and standard deviation of the proficiency variable in the no-decline class at the academic track were constrained to 0 and 1, respectively. The item position of PD onset i_0 for the 2PDM was defined by means of model fit comparisons: the model was repeatedly estimated with switching points $\delta_{i_0} = 10, \ldots, 23$, and the switching point was chosen where the BIC value was lowest. Based on the BIC, the model with a PD onset $i_0 = 18$ showed the best fit to the data and was then used for the further analyses.

The models were estimated by means of marginal maximum likelihood estimation using an expectation-maximization algorithm and using numerical integration with 15 integration points in Mplus 7.4 (L. K. Muthén & Muthén, 1998-2012). One problem with maximum likelihood estimation for mixture IRT models is that the solution can converge to a local rather than the global maximum (Finch & French, 2012). Therefore, usage of multiple random starting values is recommended to ensure replication of the best likelihood value (Lubke & Muthén, 2005). The Mplus code for the estimated models is given in the online supplement to this article.[21]

Results

We first present the goodness-of-fit statistics for the four models (Q1) and show how accounting for PD impacted item parameter estimation (Q6). Next, we present the com-

[21] URL: https://static-content.springer.com/esm/art%3A10.1186%2Fs40536-017-0049-3/MediaObjects/40536_2017_49_MOESM1_ESM.pdf

parisons between students at the academic track (*academic group* [*aca*]) and students at the non-academic school tracks (*non-academic group* [*nac*]) across models; these results are divided into three subsections. In the first subsection, we examine school-type differences in PD model parameters (Q2 to Q4). The next subsection discusses our findings on the relationship between PD and proficiency (Q5) and, in the last subsection, we present the results concerning the impact of PD on proficiency estimates (Q5a, Q8). While presenting the results, we also highlight the similarities and differences between the results provided by the different mixture PD models (Q7).

Model Fit and Item Parameter Estimates Across Models

Model fit. The model fit indices (AIC, BIC) for the different mixture PD models as well as for the 2PLM are presented in Table 5.1. The MPDM had the best fit to the data, but all mixture PD models were better than the 2PLM. Furthermore, the fit indices of the HYBRID appeared to be closer to the MPDM than to the 2PDM.

Table 5.1: Model Fit

Model	No. of free parameters	LL	AIC	BIC
2PLM	63	-217085	434296	434763
2PDM	91	-214224	428630	429304
HYBRID	71	-213978	428098	428624
MPDM	71	-213964	428070	428596

Note: LL: Log likelihood; AIC (BIC): Akaike's (Bayesian) Information Criterion; 2PLM: multigroup 2PLM; 2PDM: multigroup two-class performance decline model; HYBRID: multigroup HYBRID model; MPDM: multigroup multiclass performance decline model.

Item parameter estimates for the no-decline class. Accounting for PD is expected to reduce bias in the parameter estimates (intercepts and slope) of items affected by PD (Suh et al., 2012; Wollack et al., 2003). Since the true item parameters were unknown,

we were not able to evaluate whether any of the mixture PD models were able to do this. However, by comparing the item parameters of the no-decline class across models, we were able to investigate whether the estimates of the mixture PD models differed from the 2PLM as well as from one another.

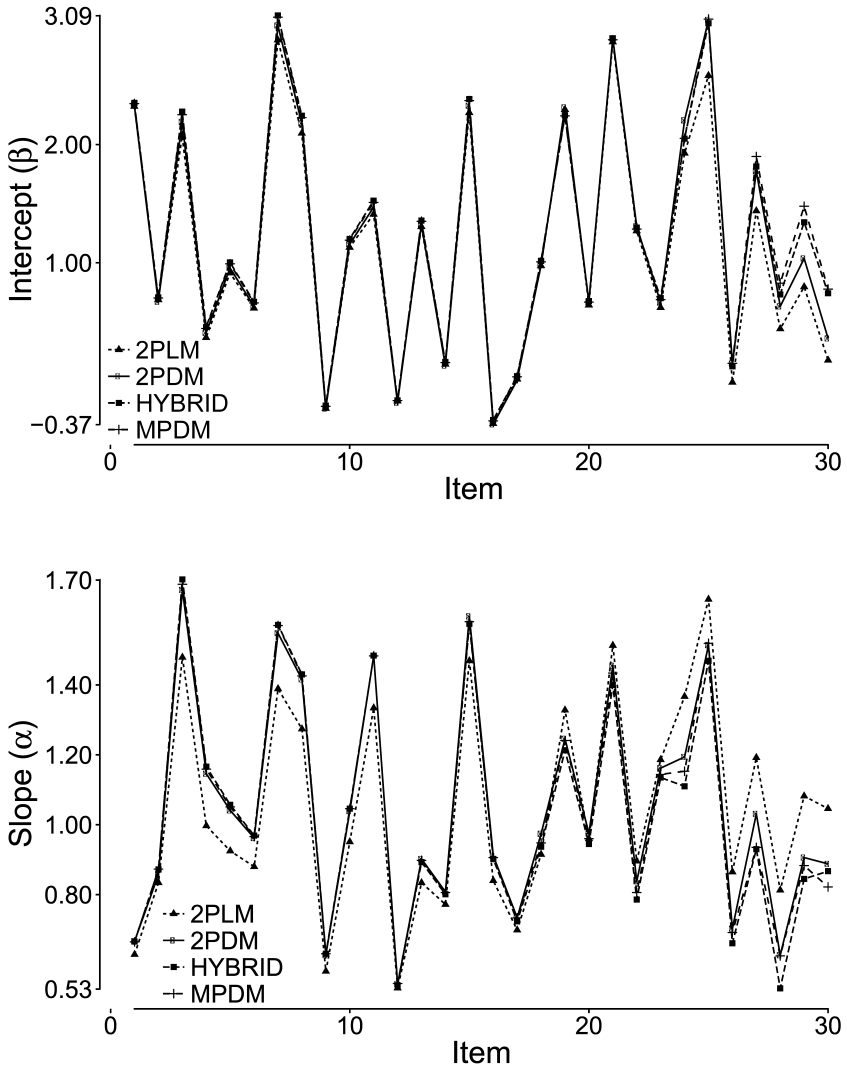

Figure 5.1: Estimates of item parameters across models. For the mixture PD models, the displayed item parameters are those of the no-decline class. 2PLM = multi-group 2PL model; 2PDM = multi-group two-class performance decline model; HYBRID = multi-group HYBRID model; MPDM = multi-group multi-class performance decline model.

In Figure 5.1 the estimated item parameters of the no-decline class (intercept, slope) for all models are depicted. For the intercept parameters, we found differences between the 2PLM and the mixture PD models for the end-of-test items. All of the mixture PD models showed higher estimates for the intercept parameters of the end-of-test items than the 2PLM, which means that items appeared to be less difficult than in the 2PLM. The mixture PD models also differed from one another: while the HYBRID and the MPDM appeared to have rather similar parameter estimates, the intercept estimates of the 2PDM were somewhat lower and closer to the 2PLM.

The slope parameter estimates were quite similar across all of the mixture PD models. Slope estimates differed between the 2PLM and the mixture PD models, not only for the end-of-test items but also for the items at the beginning of the test: while, early in the test, the estimates for the decline models were higher than those for the 2PLM, they were lower at the end of the test. These results indicate that, in the 2PLM, end-of-test items appeared to be more discriminating, a result also reported by Suh et al. (2012). However, in our application, items positioned early in the test appeared to be less discriminating than in the mixture PD models.

Group Differences in Performance Decline

The main results of all of the group comparisons regarding PD as well as proficiency are displayed in Table 5.2.

Proportions of no-decline classes and distribution of decline class proportions. The proportions of the no-decline class within each group (i.e. π_{Ig}) are displayed in Table 5.2, in the section entitled *Distribution of latent classes*. The size of π_{Ig} indicates the number of students who did not show PD during the test. The smaller the π_{Ig}, the smaller

the number of students who did not show PD was and, thus, the higher the proportion of students showing PD in the sample was.

Consistently, in all mixture PD models, π_{Ig} was significantly lower in the non-academic group (Wald's χ^2-test statistics in Table 5.2), indicating that more students at the non-academic tracks showed PD. However, the no-decline class proportions varied across models. In the 2PDM, the proportions of the no-decline classes were highest: $\pi_{I,\text{aca}} = .91$ versus $\pi_{I,\text{nac}} = .81$ (see Table 5.2, Section *Proportion of no-decline class*). As a consequence, the total number of students showing PD was smallest for the 2PDM. In contrast, the proportions of the no-decline class were smallest for the MPDM: here, the no-decline class proportion in the academic group was $\pi_{I,\text{aca}} = .68$, which means that one third of the students showed PD. In the non-academic group, the no-decline class proportion was significantly lower, with $\pi_{I,\text{nac}} = .47$. Thus, one half of the students in the non-academic group showed PD.

While, for the 2PDM, there was only one decline class for each group, for the HYBRID and the MPDM, there were multiple decline classes. As displayed in Equation 5.2, the distribution of decline classes was defined by the no-decline class proportion, π_{Ig}, and the shape parameter, ω_g: the higher ω_g, the steeper the increase in decline class proportions across item positions; yet, the higher the π_{Ig}, the smaller the cumulated probability of the decline classes and, hence, the smaller the total number of students showing PD. The cumulated class probabilities across item positions for both models are displayed in Figure 5.2. The values for ω_g are displayed in Table 5.2 (*Shape parameter*).

For the HYBRID, ω_g was significantly higher in the academic group, $\omega_{\text{aca}} = 7.33$ versus $\omega_{\text{nac}} = 4.78$. Thus, the increase in decline class proportions toward the end of the test was steeper in the academic group (Figure 5.2). Similar results were obtained for the MPDM: likewise, ω_g was significantly higher in the academic group ($\omega_{\text{aca}} = 10.27$ vs. $\omega_{\text{nac}} = 6.47$,

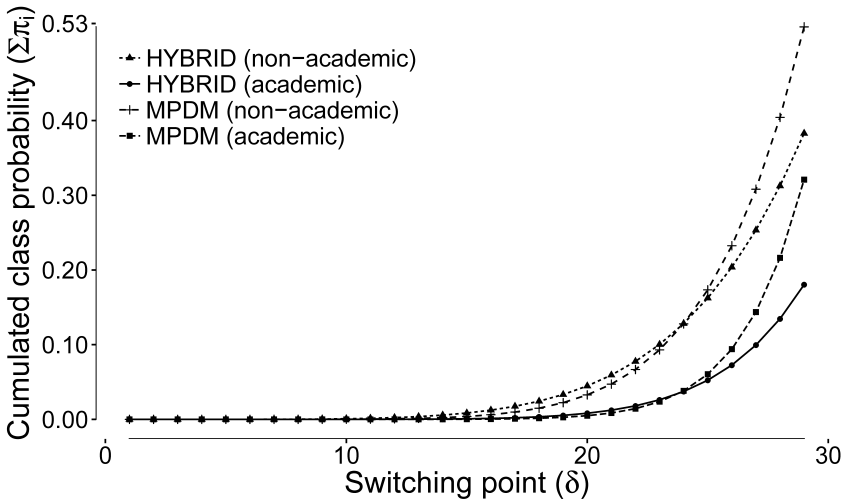

Figure 5.2: Cumulated decline class probabilities across item positions for multi-class mixture PD models. HYBRID = multi-group HYBRID model; MPDM = multi-group multi-class performance decline model.

Table 5.2), which means that the increase in class probabilities in the academic group was steeper than in the non-academic group (Figure 5.2).

Thus, the HYBRID and the MPDM provided a qualitatively similar distribution of PD classes and group differences within those classes. In both models, the no-decline class was larger in the academic track ($\pi_{I,\text{aca}} > \pi_{I,\text{nac}}$) and the majority of academic track students showing PD experienced PD onset later in the test ($\omega_{\text{aca}} > \omega_{\text{nac}}$). However, in the MPDM, class proportions $\pi_{I,g}$ were smaller and shape parameters ω_g were higher in both groups, leading to a steeper increase in cumulated class probabilities for the last five item positions (i.e., for switching points $\delta \geq 24$; Figure 5.2). However, in earlier positions, the cumulated class probabilities for both models appeared to be almost identical (see Figure 5.2). Thus, differences between the decline class proportions of the two models were mainly found for PD onset in the last five items.

Magnitude of performance decline. Group differences in the magnitude of PD were evaluated separately for each mixture PD model according to the respective model pa-

rameters assessing the magnitude of PD. In the 2PDM, the group differences found in the item intercept parameters after the switching point (i.e., $\tilde{\beta}_{ig}$, Equation 5.4) were regarded as a measure of differential PD effects, while the differences in item intercept parameters between classes affected by PD and no-decline classes (i.e., β_i vs. $\tilde{\beta}_{ig}$, Equation 5.4) were regarded as a measure of the magnitude of PD. In Figure 5.3, the intercept parameters for the no-decline class as well as for the decline classes for both groups are displayed. The intercept parameter estimates were lower in the decline classes and, thus, the probability of giving a correct response to items after the switching point was lower for students showing PD. Moreover, the intercept parameters decreased across item positions, so that items appeared to become gradually more difficult toward the end of the test in both groups. Furthermore, the results displayed in Figure 5.2 do not provide a sound indication of group differences in the onset point of PD as the estimates of the $\tilde{\beta}_{\text{ig}}$-paramters decreased immediately after the onset point in both groups. However, the intercept parameters after the switching point were significantly lower in the non-academic group (Wald's $\chi^2\,(12) = 25.24$, $p = 0.01$), which means that the average size of the PD appeared to be larger in the non-academic school tracks. However, compared to the large difference in intercept parameters between the no-decline and the decline classes, the group differences appeared rather small.

In the HYBRID, group differences in the magnitude of PD were reflected in differences in the response threshold $\tilde{\beta}_g$ (Equation 5.5). The response thresholds were $\tilde{\beta}_{\text{nac}} = -3.76$ and $\tilde{\beta}_{\text{aca}} = -3.48$ for the non-academic and academic group, respectively, which did not differ significantly from one another (see Table 5.2, Section *Magnitude of PD*). After the switching point, the probability of a correct response being provided was .02 in the non-academic group and of .03 in the academic group of obtaining a correct response after the switching point.

In the MPDM, the probability of obtaining a correct response was not constant after the switching point. Rather, in each decline class, this probability decreased, depending on

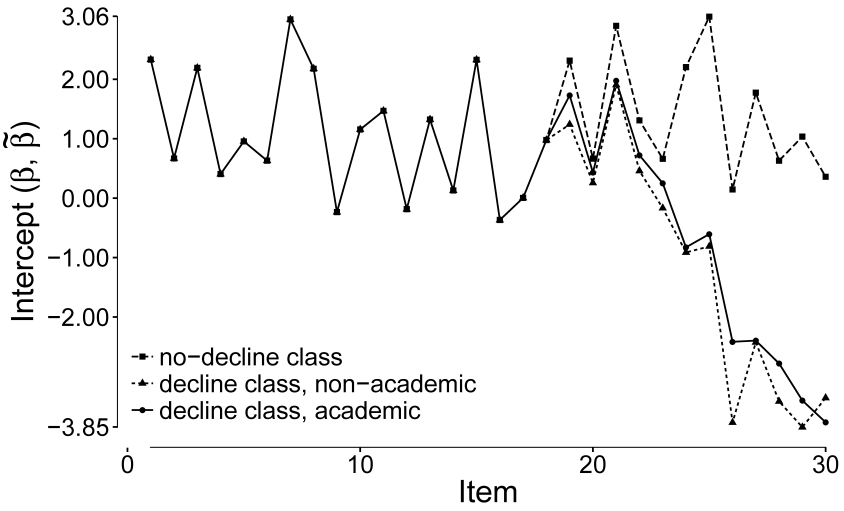

Figure 5.3: Intercept parameters of no-decline and decline classes (multigroup two-class performance decline model). Item parameters of the no-decline classes are invariant between groups.

the PD onset point, with group-specific decrement parameters κ_g (Equation 5.7). However, κ_g did not differ significantly between groups ($\kappa_{\text{nac}} = 0.80$ vs. $\kappa_{\text{aca}} = 0.91$, see Table 5.2, Section *Magnitude of PD*), indicating that the magnitude of PD depended only on the PD onset point, and not on the school track the students attended.

Hence, it can be summarized that, of the mixture PD models, the HYBRID and the MPDM provided no indication that the magnitude of PD differs between school tracks. In contrast, the 2PDM revealed statistically significant group differences in the magnitude of PD, with stronger declines in non-academic track students. However, from a substantive point of view, the difference was rather small.

Relationship Between Performance Decline and Proficiency

In the 2PDM for both groups, the mean of θ appeared to be somewhat higher in the decline class. In the non-academic group, the mean proficiency was $\mu_{\theta,I,nac} = -1.02$ in the no-decline class and $\mu_{\theta,i_0,nac} = -0.95$ in the decline class, but the difference was not statistically significant at the $p \leq .05$ level (Table 5.2, Section *No-decline class: mean*). In

the academic group, the mean and standard deviation in the no-decline class were fixed at 0 and 1, respectively. The mean proficiency in the decline class was $\mu_{\theta,i_0,aca} = 0.03$ which did not differ significantly from the no-decline class (Table 5.2).

For the HYBRID and the MPDM, the mean θ followed a linear function over δ with group-specific slopes ρ_g (Equation 5.6). For the HYBRID, in both groups, ρ_g was negative. This indicates that the mean of θ was lower in the decline classes than in the no-decline class and that it further decreased when PD onset occurred in earlier item positions. Furthermore, ρ_g was significantly lower in the academic group ($\rho_{\text{aca}} = -0.05$ vs. $\rho_{\text{nac}} = -0.01$, Table 5.2, Section *Decline classes*), indicating a stronger negative association between θ and PD in the academic group. More specifically, the value of the ρ_{aca}-parameter of -0.05 means that students who had a PD onset at item 25 (i.e., 5 items before the end of the test) were expected to score 0.25 units lower on the proficiency variable. As the standard deviation of θ was fixed to one in the no-decline class in the academic track, this result directly reflects a standardized effect size.

The results for the MPDM were quite similar to those for the HYBRID. In both groups, ρ_g was also negative and, likewise, the slope in the academic group was significantly lower ($\rho_{\text{aca}} = -0.03$ vs. $\rho_{\text{nac}} = -0.01$, Table 5.2). Thus, the average proficiency was lower in the decline classes and the decrease in proficiency when PD onset occurred earlier in the test was steeper in the academic group.

Hence, the HYBRID and the MPDM converged in their conclusion about the relationships between students' proficiencies and their PD onset points, as both models indicated that the earlier the students' PD onset occurred, the lower their proficiency was, and, in both models, this relationship was more exaggerated in academic track students. In contrast, the 2PDM provided no indication of a relationship between proficiency and PD.

Impact of Accounting for Performance Decline on Proficiency Estimation

Comparisons of individual proficiency scores. In order to illustrate how accounting for PD affects the proficiency distribution, we compared the expected a posteriori (EAP) scores for each model between groups and PD classifications (i.e., students not showing PD [*no-PD class*] vs. students showing PD [*PD class*]).[22] In order to be able to compare score estimates, we transformed the EAP scores obtained by the PD models to the metric of the 2PLM by means of linear equating (e.g., Livingston, 2014).

In Figure 5.4, the equated EAP scores for each mixture PD model are plotted against the EAP scores estimated by the 2PLM, displayed separately for school tracks and PD classification. Across all mixture PD models and across both groups, for the no-PD classes, the EAP score estimates of the mixture PD models appeared to be almost identical to those of the 2PLM. However, for students belonging to the PD classes, the EAP scores appeared to be higher when estimated by one of the mixture PD models than those estimated by the 2PLM model. Assuming that scores estimated by the mixture PD models are more accurate reflections of proficiency, the EAP scores estimated by the 2PLM appeared to underestimate the proficiency of test takers showing PD. Looking at the score estimates for the PD classes across models, the correspondence between the EAP scores of the 2PLM and the 2PDM was very high because, in the 2PDM, only one decline class was estimated for each group. For both the HYBRID and the MPDM, the difference in EAP scores between the mixture PD model and the 2PLM depended on the switching point and, thus, the correspondence between scores was lower (Figure 5.4). Interestingly, the EAP scores

[22] For PD classification purposes, the probability of the no-decline class was dummy coded so that students were classified as belonging to the *no-PD class* if their probability of belonging to the no-decline class was larger than .5. Otherwise, they were classified as belonging to the combined *PD class*.

derived from the HYBRID and the MPDM showed very high agreement, although they were based on models that differed in the specification of PD.

Group differences in proficiency. We also investigated whether accounting for PD would affect the estimation of group differences in proficiency. In order to compare the magnitude of group differences as estimated by the four models, we calculated effect sizes between the group means of proficiency. For the mixture PD models, the group means were calculated from the marginal distributions of θ across latent classes. [23] The group means and standard deviations of θ and the effect sizes (d) for all models are displayed in Table 5.3.

Across all models, the average proficiency in the non-academic group was lower than in the academic group, as expected. In all of the mixture PD models, the effect sizes of the group difference appeared to be very similar to one another (about $d = 1.09$) and smaller than for the 2PLM ($d = 1.17$; Table 5.3). Thus, groups differed to a smaller degree when mixture PD models were employed, regardless of the specification of PD. However, in the present application, proficiency differences between groups were about one standard deviation for all models, which means that the differences in the results provided by the 2PLM and the mixture PD models are rather small when considered on a relative scale.

Summary and Discussion

The aim of the present article was to compare the potential of three mixture PD models, the 2PDM (Bolt et al., 2002), the HYBRID (Yamamoto, 1995), and the MPDM (Jin & Wang, 2014), for assessing PD in proficiency tests administered in LSAs. The models were extended to accommodate multiple groups, thereby making it possible to examine and

[23] The resulting distribution is a normal mixture distribution with a group-specific mean $\mu_g = \sum_{i=1}^{I} (\pi_{ig} \cdot \mu_{ig})$ and a standard deviation as $\sigma_g = \sqrt{\left(\sum_{i=1}^{I} \left[\pi_{\text{ig}} \cdot \left(\mu_{\text{ig}}^2 + \sigma_g^2\right)\right]\right) - \mu_g}$ with $i = 1, \ldots, I$.

test group differences in PD, and to adjust group differences in proficiencies for PD. In order to examine the similarities and discrepancies in the results and conclusions provided by the multigroup mixture PD models, the models were applied to a mathematics tests administered in a German LSA. The results indicate that the models provided similar conclusions about key aspects of PD, but differed with respect to some results. However, the main results gathered by the mixture PD models were in line with results obtained in other settings (e.g., G. Nagy, Lüdtke & Köller, 2016).

The mixture PD models consistently indicated that the mathematics test was affected by PD, as all of the mixture PD models fitted the data better than a standard 2PLM. Thus, it seems reasonable to assume that, for a subsample of test takers, the responses obtained for end-of-test items reflected PD. In addition, in all of the mixture PD models, the decline class proportions were higher in the non-academic group, but the estimates of the magnitude of PD, such as the intercept parameters of items affected by PD (2PDM), the response thresholds (HYBRID), and the decrement parameters (MPDM) appeared to be similar between groups. Thus, all models indicated that groups differed mainly in the proportions of students showing PD and not in the magnitude of PD. Furthermore, in all models, group differences in the proportions of students showing PD affected the results of the group differences in proficiency. Interestingly, all mixture PD models adjusted the group differences provided by the 2PLM to a similar extent. However, the reductions in the effect sizes were not large when considered on a relative scale; this was due to the very strong track differences in proficiency. Nevertheless, in other settings, where group differences in proficiencies are smaller, the adjustments provided by the mixture PD models might lead to qualitatively different conclusions.

Additionally, all mixture PD models provided relatively similar estimates of item parameters that were different from the parameters estimated by the standard 2PLM. As expected, end-of-test items, that is, those items that were most strongly affected by PD,

were estimated to be less difficult than in the 2PLM, and the slope parameters belonging to these items were estimated to be lower by the mixture PD models as compared to the 2PLM. These results are in line with the simulation study of Suh et al. (2012). However, in contrast to Suh et al. (2012), we found that the mixture PD models provided higher slope parameters for items positioned at the beginning of the test. One reasonable explanation for this result is that individual differences in PD onset points not only increased the dependencies between items at the end of the test, but also reduced the relationships between items at the beginning of test. Hence, when PD was not accounted for, the slope parameters of the 2PLM were estimated to account for the strong dependencies between end-of-test items (i.e., higher slopes for items affected by PD) and the weaker associations between items located at the beginning and the end of the test (i.e., lower slopes for items not affected by PD). Of course, more research is clearly needed to examine the plausibility of our explanation.

Despite the similarities, the three mixture PD models differed from each other in some respects. All of the mixture PD models differed in their estimation of the number of students showing PD: the decline class proportions were smallest in the 2PDM and largest in the MPDM. However, the differences between the MPDM and the HYBRID were rather subtle because the distribution of PD onset points differed only with respect to the last five items. In addition, the mixture PD models provided different conclusions about the relationships between PD onset points and proficiency. In the 2PDM, there were no significant proficiency differences between decline and no-decline classes, neither for students at the academic nor for students at the non-academic school tracks. For both the HYBRID and the MPDM, the mean proficiency was lower in the decline classes and it decreased when PD onset occurred nearer the beginning of the test. Furthermore, in both cases, this decrease was stronger in the academic group. Similarly, the mixture PD models differed in the adjustments of proficiency scores relative to the 2PLM. As we have shown for the EAP

scores, for students showing PD, the individual proficiency scores were estimated to be higher in the mixture PD models than in the 2PLM, but the EAP scores provided by the 2PDM were closer to those estimated by the 2PLM. Here, the HYBRID and the MPDM provided EAP scores that were quite similar to each other.

Conclusions and Future Directions

Taken together, the HYBRID and the MPDM that both consist of many latent classes performed rather similar in many respects, although they differ in their representation of PD. Hence, the number of latent classes combined with the assumptions about their distribution appears to be the main divider between the mixture PD models envisaged in this article. The mixture PD models are based on different assumptions on how PD affects test-taking behavior. In the 2PDM, it is assumed that the switching point is identical for all test takers showing PD within a group. In contrast, in both the HYBRID and the MPDM, multiple switching points are considered, but the models differ in their assumptions about PD. In the HYBRID, it is assumed that the probability of a correct response occurring after the switching point no longer depends on proficiency, whereas, in the MPDM, it is assumed that responses given after the switching point still depend on proficiency. This assumption is also embedded in the 2PDM. However, the question of whether the specification of PD provided by different mixture models are of less importance than the assumptions about the number of PD classes remains open, as suggested by the results provided in this article. Therefore, further research on this issue is called for.

In further applications, the suitability of the proposed model restrictions should be analyzed thoroughly and adapted if necessary. For the HYBRID and the MPDM, we assumed that there was a linear relationship between proficiency and switching points. For some applications, this restriction might be too strict, for example, the decrease in proficiency could be stronger for a switching point in earlier item positions and could

diminish when PD affects only the very last items. Moreover, the distribution of decline classes could be modeled by functions other than the one proposed in our study. The issue of model restrictions warrants consideration in subsequent research.

In our study, the HYBRID and the MPDM lead to similar results. Thus, both models appear to be comparably well suited to explore PD in educational LSAs. However, more research is needed on the similarities between these two models.

Finally, the variations of PD effects across domains should be explored. Research on item position effects shows that the strength of the effects varies with respect to the domain being studied (Debeer & Janssen, 2013; G. Nagy, Lüdtke & Köller, 2016). Comparing PD in several domains for the same population could also be useful in determining whether PD can be regarded as an overarching person characteristic or, rather, as a test-specific phenomenon. Mixture PD models with multiple decline classes provide estimates of the switching point and its covariation with proficiency, and this allows for fine-grained analyses of PD. In this sense, mixture PD models can be used to study how PD interacts with proficiency and other variables to provide a better understanding of the mechanisms behind PD.

Table 5.2: Parameter estimates of performance decline and proficiency

	nac Est. (SE)	aca Est. (SE)	Group comparisons Wald's χ^2(df)
Distribution of latent classes			
Proportion of no-decline class (π_I)			
2PDM	0.81 (0.01)	0.91 (0.01)	106.80 (1), p<0.001
HYBRID	0.62 (0.01)	0.82 (0.01)	226.86 (1), p<0.001
MPDM	0.48 (0.03)	0.68 (0.03)	41.60 (1), p<0.001
Shape parameter (ω)			
2PDM			
HYBRID	4.78 (0.15)	7.33 (0.27)	72.27 (1), p<0.001
MPDM	6.47 (0.26)	10.25 (0.47)	70.52 (1), p<0.001
Magnitude of PD			
2PDM	_[a]	_[a]	_[a]
HYBRID: response threshold ($\tilde{\beta}$)	3.76 (0.18)	3.48 (0.31)	0.65 (1), p = 0.42
MPDM: decrement (κ)	0.80 (0.04)	0.91 (0.07)	2.85 (1), p = 0.09
Proficiency distribution			
No-decline class: mean (μ_I)			
2PDM	-1.02 (0.03)	0b	1637.64 (1), p<0.001
HYBRID	-1.01 (0.03)	0b	1385.61 (1), p<0.001
MPDM	-1.01 (0.03)	0b	1321.48 (1), p<0.001
No-decline class: standard deviation (σ)			
2PDM	0.82 (0.02)	1[b]	133.48 (1), p<0.001
HYBRID	0.82 (0.02)	1[b]	124.16 (1), p<0.001
MPDM	0.82 (0.02)	1[b]	127.80 (1), p<0.001
Decline classes			
2PDM: mean (μ_{i_0})	-0.95 (0.04)	0.02 (0.08)	167.28 (1), p<0.001
HYBRID: slope (ρ)	-0.01 (0.004)	-0.05 (0.01)	17.29 (1), p<0.001
MPDM: slope (ρ)	-0.01 (0.004)	-0.03 (0.01)	5.66 (1), p = 0.02

Note: nac: non-academic group; aca: academic group; 2PDM: multigroup two-class performance decline model; HYBRID: multigroup HYBRID model; MPDM: multigroup multiclass performance decline model. All parameter estimates are group-specific.
[a] see Figure 5.3. [b] parameters are fixed.
Note that parameters cannot be compared across models. See main text for more information.

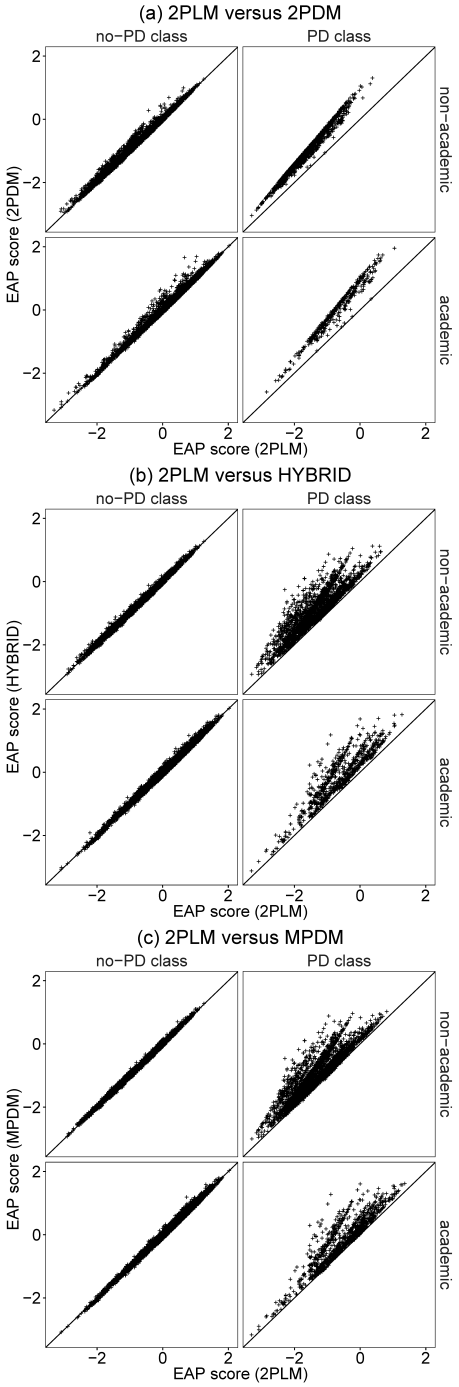

Figure 5.4: EAP scores estimated by the 2PLM versus mixture PD models by group and PD classification. x-Axis: EAP scores estimated by the multigroup 2PLM; y-Axis: equated EAP scores estimated by the multigroup two-class performance decline model (2PDM, a), by the multigroup HYBRID model (b), and by the multigroup multiclass performance decline model (MPDM, c).

Table 5.3: Group differences in mean proficiency

Model	μ_{nac}	σ_{nac}	μ_{aca}	σ_{aca}	$d(SE)$
2PLM	-1.07	0.82	0^a	1^a	1.17 (0.020)
2PDM	-1.03	0.82	-0.04	1.01	1.08 (0.020)
HYBRID	-1.00	0.82	0.00	1.00	1.10 (0.020)
MPDM	-1.03	0.82	-0.03	1.00	1.09 (0.020)

Note: nac: non-academic group; aca: academic group; 2PLM: multigroup 2PLM; 2PDM: multigroup two-class performance decline model; HYBRID: multigroup HYBRID model; MPDM: multigroup multiclass performance decline model. μ and σ obtained from the marginal distribution of θ within groups.
[a] parameters are fixed.

Kapitel 6

Gesamtdiskussion

6.1 Einleitung und Übersicht

In nahezu allen LSAs bearbeiten nicht alle Testteilnehmenden den Test bis zum Ende mit maximaler Leistung. Testteilnehmende können den Test abbrechen, was sich in der Anzahl der Not-Reached-Items zeigt, oder sie können einen Leistungsabfall zeigen, sodass die Lösungswahrscheinlichkeit am Testende sinkt. Beide Phänomene stellen eine Veränderung des Testbearbeitungsverhaltens dar. In dieser Dissertation wurden Modellierungsansätze vorgeschlagen, um die Veränderung des Testbearbeitungsverhaltens in einem Leistungstest zu erfassen und den Zusammenhang mit der Personenfähigkeit und Gruppierungsvariablen zu schätzen. Ziel war es, bestehende Modelle zu erweitern, um restriktive Annahmen zu lockern und um die Schätzung differentieller Effekte, das heißt Gruppenunterschiede in der Anzahl der Not-Reached-Items beziehungsweise im Ausmaß des Leistungsabfalls, zu ermöglichen. Im Vordergrund stand die Bewertung der Performanz der Modelle anhand empirischer Datensätze.

Im Rahmen der Dissertation wurden zwei Studien durchgeführt, die sich jeweils mit einem der beiden Aspekte der Veränderung des Testbearbeitungsverhaltens (Not-Reached-Items oder Leistungsabfall) beschäftigen. In den folgenden Abschnitten dieses Kapitels werden die Ergebnisse der beiden Studien zusammengefasst und vor dem Hintergrund des aktuellen Forschungsstands bewertet (Abschnitt 6.2), Anwendungsmöglichkeiten für die vorgestellten Modelle in LSAs diskutiert (Abschnitt 6.3) und die Grenzen der Arbeit bewertet (Abschnitt 6.4). Das Kapitel schließt mit einem Ausblick auf zukünftige Forschungsfragen (Abschnitt 6.5) und einem Resümee.

6.2 Zusammenfassung und Diskussion der Studien

6.2.1 Studie 1: Modellierung des Testabbruchs mittels Survivalanalysen

Die erste Studie der vorliegenden Arbeit beschäftigt sich mit Not-Reached-Items. Not-Reached-Items implizieren einen Testabbruch (z. B. aufgrund von Zeitdruck oder geringer Testbearbeitungsmotivation), der eine extreme Form der Veränderung des Testbearbeitungsverhaltens darstellt. In dieser Studie wurde ein Modell zum Testabbruch entwickelt, das eine flexible Modellierung des Zusammenhangs mit der Personenfähigkeit sowie weiterer Kovariaten ermöglicht. Das Modell basiert auf dem Ansatz von Glas und Pimentel (2008), die ein zweidimensionales IRT-Modell vorgeschlagen haben, das neben der Personenfähigkeit eine latente Missingness-Variable, definiert über die Not-Reached-Items, erfasst. Mit der Missingness-Variable wird die Itemposition kodiert, ab der eine Person die Testbearbeitung abbricht. Besteht ein signifikanter Zusammenhang zwischen der Personenfähigkeit und der Missingness-Variable, sind Not-Reached-Items als MNAR zu bewerten (vgl. Little & Rubin, 2002; Schafer & Graham, 2002) und sollten deshalb bei der Parameterschätzung berücksichtigt werden.

Als flexiblere Alternative zum Modell von Glas und Pimentel (2008) wurde in Studie 1 das *Mixture Discrete (Item) Sequence Event Model* (MDSEM) entwickelt. Das MDSEM dient zur Untersuchung von Prädiktoren des Testabbruchs. Dazu wird im MDSEM die Missingness-Variable als latente Variable mit einer beliebigen Verteilung modelliert, deren Häufigkeitsverteilung anhand von diskreten Stützstellen geschätzt wird. Dies wird über ein *located latent class model* (Formann, 1985) realisiert, bei dem die Klassenvariable die Stützstellen der Häufigkeitsverteilung der Missingness-Variable kodiert und die Klassengrößen der jeweiligen Häufigkeit an dem Punkt entsprechen. Mittels einer multinomialen Regression auf die Fähigkeitsvariable wird ein nichtlinearer Zusammenhang zwischen bei-

den Variablen geschätzt. Anders als im Modell von Glas und Pimentel (2008) gibt es im MDSEM keine Restriktionen für die Wahrscheinlichkeiten für einen Testabbruch im Testverlauf.

Während im Modell von Glas und Pimentel (2008) ein korrelativer Zusammenhang zwischen Personenfähigkeit und Missingness-Variable betrachtet wird, wird im MDSEM ein gerichteter Zusammenhang geschätzt, wobei die Missingness-Variable die abhängige Variable darstellt. Personenfähigkeit und weitere Variablen werden als Prädiktoren für den Testabbruch modelliert. Der Einbezug weiterer Kovariaten erlaubt die Untersuchung differentieller Effekte bei gleichzeitiger Kontrolle der Personenfähigkeit. Das MDSEM hat wie das Modell von Glas und Pimentel (2008) Ähnlichkeiten mit einer Survivalanalyse, in der die Wahrscheinlichkeit, an einer bestimmten Itemposition den Test abzubrechen, in Abhängigkeit von Kovariaten modelliert wird. Die Survivalanalyse bietet ein anschauliches Konzept für die Interpretation der Ergebnisse. Die in der Survivalanalyse üblichen grafischen Darstellungen (z. B. die Survivalfunktion) ermöglichen, auch komplexe Beziehungen in einfacher Weise zu veranschaulichen, was die Interpretation der Ergebnisse in der praktischen Anwendung erleichtert.

6.2.1.1 Zusammenfassung der Ergebnisse

Die Performanz des MDSEM wurde in zwei empirischen Datensätzen zu Mathematikleistungstests und in einer Simulationsstudie untersucht. Das MDSEM wurde auf eine Stichprobe von Schülerinnen und Schülern (SuS) der siebten Klasse und eine Stichprobe von Auszubildenden im ersten Ausbildungsjahr angewandt und mit dem Modell von (Glas & Pimentel, 2008) verglichen. Es zeigt sich in beiden Datensätzen ein besserer Modellfit für das MDSEM. In beiden Stichproben implizieren das MDSEM wie auch das Modell von Glas und Pimentel (2008) einen negativen Zusammenhang[24] zwischen Mathematikfähig-

[24]Durch die Kodierung der Missingness-Variable bedeuten höhere Werte *weniger* Not-Reached-Items, also einen späteren Testabbruch.

keit und der Missingness-Variable, das heißt Personen mit geringerer Fähigkeit zeigen einen späteren Zeitpunkt des Testabbruchs.

Wird das MDSEM um die Gruppenzugehörigkeit als Kovariate erweitert, finden sich in beiden Datensätzen differentielle Effekte: Fähigkeit und Testabbruch sind innerhalb der Gruppen weiterhin negativ assoziiert, aber Personen in einer technischen Ausbildung und Schülerinnen und Schüler an nichtgymnasialen Schulformen zeigen tendenziell einen früheren Testabbruch als Personen in einer industriekaufmännischen Ausbildung beziehungsweise Schülerinnen und Schülern am Gymnasium mit vergleichbarer Personenfähigkeit. Die Wahrscheinlichkeiten für den Testabbruch sind für Personen mit mittleren und höheren Personenfähigkeiten dabei noch einmal stärker ausgeprägt, es zeigen sich also nichtlineare Effekte in beiden Stichproben.

In einer Simulationsstudie wurde untersucht, wie sich die Berücksichtigung der Missingness-Variable auf die Schätzung des Gruppenunterschieds in der Fähigkeit und die Schätzung der Itemparameter auswirkt. Hängt der Testabbruch mit der Personenfähigkeit zusammen, zeigt sich, dass der Fähigkeitsunterschied zwischen den Gruppen unterschätzt wird, wenn der Zusammenhang in der Parameterschätzung vernachlässigt wird. Auch die Varianz der Personenfähigkeit wird dann unterschätzt, während die Itemparameter kaum verzerrt geschätzt werden. Im Gegensatz dazu liefert das MDSEM nahezu unverzerrte Schätzungen.

Diese Studie präsentiert mit dem MDSEM eine Erweiterung des Modells von Glas und Pimentel (2008), die die Modellierung von flexiblen Zusammenhängen zwischen Testabbruch und Fähigkeit ermöglicht. Die Ergebnisse der Simulationsstudie stützen die bisherigen Befunde zur Verzerrung der Item- und Personenparameter und zeigen darüber hinaus, dass das Ignorieren der Not-Reached-Items zu einer Verzerrung der Gruppenunterschiede in der Personenfähigkeit führen kann.

6.2.1.2 Einordnung der Befunde

Die meisten Modelle für fehlende Itemantworten unter MNAR-Bedingungen wurden für ausgelassene Items entwickelt und legen den Fokus auf die Schätzung der Personenfähigkeit nach Kontrolle des Einflusses der Missingness-Variable, um den Bias zu reduzieren (z. B. Glas et al., 2015; Holman & Glas, 2005; Rose et al., 2015, 2017). Für Not-Reached-Items haben Glas und Pimentel (2008) ein Modell vorgeschlagen, das ebenfalls die Reduktion von Bias in der Parameterschätzung zum Ziel hat. Das MDSEM unterscheidet sich von den bisherigen Ansätzen, weil es einerseits fehlende Antworten auf Not-Reached-Items und nicht ausgelassene Items betrachtet und andererseits den Einfluss von Personenfähigkeit und weiterer Kovariaten auf die Missingness-Variable als abhängige Variable untersucht. Für die Missingness-Variable wird keine Normalverteilung angenommen und es kann ein komplexer, nichtlinearer Zusammenhang mit den Prädiktoren des Testabbruchs geschätzt werden. Diesen Ansatz, die Missingness-Variable anhand einer potentiell nichtnormalverteilten, diskreten Verteilung darzustellen, hat das MDSEM mit dem Modell von Köhler et al. (2015b) gemeinsam. Im Gegensatz zum MDSEM werden in diesem Modell ausgelassene statt Not-Reached-Items behandelt und die Verteilung der Missingness-Variable wird anders spezifiziert. Zusätzlich wird auch die Fähigkeitsvariable durch eine diskrete Verteilung dargestellt.

In beiden Stichproben von Studie 1 zeigte sich, dass Personen mit hoher Mathematikfähigkeit eine höhere Wahrscheinlichkeit haben, den Test früh abzubrechen. Diese Befunde stehen im Gegensatz zu den Ergebnissen von Pohl et al. (2014) und Rose et al. (2010), die jeweils finden, dass Testteilnehmende mit geringer Mathematikfähigkeit mehr Not-Reached-Items haben. Bei der Studie von Rose et al. (2010) ist jedoch zu beachten, dass hier die Missingness-Variable über die Summe von ausgelassenen und Not-Reached-Items sowie ungültige Antworten definiert ist. Die Ergebnisse lassen sich damit nur bedingt mit den anderen in Beziehung setzen. Die beiden Studien untersuchen zudem andere Alters-

gruppen, Schülerinnen und Schüler der fünften Klasse (Pohl et al., 2014) beziehungsweise eine Stichprobe aus PISA 2006 (Rose et al., 2010).

Der Befund, dass Schülerinnen und Schüler an nicht-gymnasialen Schulformen einen früheren Testabbruch zeigen als Personen vergleichbarer Fähigkeit am Gymnasium, ähnelt den Ergebnissen aus der Studie von Köhler et al. (2015a). Allerdings findet sich dieser Effekt bei Köhler et al. (2015a) nur für die Stichprobe der neunten Klasse, nicht für die fünfte Klasse. In der Studie 1 dieser Dissertation wurde eine Stichprobe an der siebten Klasse untersucht. Die Ergebnisse beider Studien können so auf die Altersabhängigkeit differentieller Effekte der Schulform hindeuten.

6.2.2 Studie 2: Modellierung des Leistungsabfalls mittels Mischverteilungsmodellen

In der zweiten Studie der vorliegenden Arbeit wurden die Mischverteilungsmodelle von Bolt et al. (2002), Jin und Wang (2014) und Yamamoto (1995) zur Modellierung von Leistungsabfall miteinander verglichen. Allen drei Modellen liegt die Annahme zugrunde, dass die Stichprobe der Testteilnehmenden aus Personen mit und ohne Leistungsabfall besteht. Personen, die einen Leistungsabfall zeigen, bearbeiten den Test bis zu einem bestimmten Schaltpunkt mit maximaler Leistung und ändern dann ihr Bearbeitungsverhalten. Ziel der Modelle ist es, die Stichprobe in latente Klassen zu zerlegen, innerhalb derer alle Personen dasselbe Ausmaß von Leistungsabfall zeigen. Die Modelle unterscheiden sich in der Operationalisierung des Leistungsabfalls und den Annahmen zum Schaltpunkt.

Bolt et al. (2002) gehen davon aus, dass alle Testteilnehmenden mit einem Leistungsabfall an derselben Itemposition ihr Bearbeitungsverhalten ändern. Damit werden zwei latente Klassen von Testteilnehmenden definiert. Der Leistungsabfall spiegelt sich höheren Itemschwierigkeiten für die Items nach dem Schaltpunkt wider. Jin und Wang (2014)

nehmen an, dass ein Schaltpunkt in der Testbearbeitung zu jedem Zeitpunkt im Test vorkommen kann. Für jede Ausprägung der Schaltpunkt-Variable (d. h. für jede Itemposition) wird eine latente Klasse definiert. Der Leistungsabfall zeigt sich in diesem Modell in einer je latenter Klasse spezifischen Reduktion der Personenfähigkeit. Auch Yamamoto (1995) nimmt an, dass der Schaltpunkt an jeder Itemposition vorkommen kann. In diesem Modell wird angenommen, dass die Itemantworten nach dem Schaltpunkt nicht mehr von der Personenfähigkeit abhängen, weil die Testteilnehmenden ab diesem Punkt unsystematisches Rateverhalten zeigen.

Die drei Mischverteilungsmodelle wurden zu Mehrgruppenmodellen erweitert, die ermöglichen, Gruppenunterschiede im Ausmaß des Leistungsabfalls zu schätzen. Die Gruppen können sich dabei einerseits in der Stärke des Leistungsabfalls und andererseits in den Größen der latenten Klassen unterscheiden – also zum einen im Ausmaß der Veränderung des Testbearbeitungsverhaltens und zum anderen in der Anzahl der Personen mit einem Leistungsabfall. Die resultierenden Mehrgruppenmodelle wurden auf einen Mathematiktest angewandt, um Schulformunterschiede im Ausmaß des Leistungsabfalls zu erfassen und um Ähnlichkeiten und Unterschiede in den Implikationen der drei Modelle zu untersuchen.

6.2.2.1 Zusammenfassung der Ergebnisse

In allen drei Modellen wird der Anteil der Personen mit Leistungsabfall in den nichtgymnasialen Schulformen höher eingeschätzt, jedoch zeigen sich keine bedeutsamen Gruppenunterschiede im Ausmaß des Leistungsabfalls. Die Modelle von Jin und Wang (2014) und Yamamoto (1995) implizierten, dass Personen ohne Leistungsabfall im Mittel eine höhere Personenfähigkeit haben und dass die Personenfähigkeit umso geringer ist, je früher im Test der Leistungsabfall einsetzt. Das Modell von Bolt et al. (2002) indiziert hingegen keinen Zusammenhang zwischen Personenfähigkeit und Leistungsabfall. Wenn für den Leistungsabfall kontrolliert wird, zeigt sich gegenüber einem Standard-IRT-Modell ei-

ne Reduzierung im Schulformunterschied in der Personenfähigkeit. In allen drei Modellen ist dabei die Reduktion sehr ähnlich. Die Itemschwierigkeiten zeigen in den Mischverteilungsmodellen die erwarteten Abweichungen gegenüber dem Standard-IRT-Modell: Die Schwierigkeiten der Items am Testende wurden in den Mischverteilungsmodellen niedriger geschätzt, was auf die Berücksichtigung der durch den Leistungsabfall verringerten Lösungswahrscheinlichkeit in den Items zurückzuführen ist.

Für diese Studie wurden die Not-Reached-Items als falsche Antwort kodiert. Die sehr hohen Schwierigkeiten der Items nach dem Schaltpunkt im Modell von Bolt et al. (2002) sowie der Parameter der Antwortschwelle im Modell von Yamamoto (1995) implizieren, dass die Lösungswahrscheinlichkeit für Items nach dem Schaltpunkt nahe Null ist. Dies deutet darauf hin, dass der Leistungsabfall in dieser Studie mit Not-Reached-Items konfundiert ist: Da der Test aus Multiple-Choice-Items besteht, wäre etwa für das Modell von Yamamoto (1995) zu erwarten, dass die Lösungswahrscheinlichkeiten für Items am Testende ungefähr der Ratewahrscheinlichkeit entsprechen.

Unterschiede zwischen den drei Modellen zeigen sich vor allem für den Anteil der Personen ohne Leistungsabfall: Im Modell von Jin und Wang (2014) zeigen über die Gesamtstichprobe hinweg etwa 60 %, im Modell von Yamamoto (1995) 70 % und im Modell von Bolt et al. (2002) etwa 85 % der Testteilnehmenden keinen Leistungsabfall. Beim Vergleich der drei Mischverteilungsmodelle zeigt sich, dass das Modell von Jin und Wang (2014) die beste und das Modell von Bolt et al. (2002) die geringste Passung auf die Daten aufweist.

Die Modellierung von mehreren Schaltpunkten, an denen ein Leistungsabfall einsetzen kann, wie im Modell von Jin und Wang (2014) und im Modell von Yamamoto (1995) vorgenommen, erscheint für die meisten Testdesigns plausibler als die Beschränkung auf einen einzigen Schaltpunkt, wie es im Modell von Bolt et al. (2002) angenommen wird. Das Modell von Jin und Wang (2014) erlaubt, unterschiedliche Startpunkte und zusätzlich ein unterschiedliches Ausmaß des Leistungsabfalls zu modellieren, was für viele Anwendun-

gen als realistische Operationalisierung von Leistungsabfall erscheint. Im Gegensatz dazu betont das Modell von Yamamoto (1995) die qualitative Veränderung des Testbearbeitungsverhaltens, was zum Beispiel in Testsituationen, wo Rateverhalten wahrscheinlich ist, von Interesse ist. Dass der Anteil der Testteilnehmenden mit Leistungsabfall im Modell von Bolt et al. (2002) vergleichsweise gering eingeschätzt wurde, kann an der zu restriktiven Annahme eines einzigen Schaltpunkts für den Leistungsabfall liegen, sodass Personen, die erst spät im Test einen Leistungsabfall zeigen, der Klasse ohne Leistungsabfall zugeordnet wurden.

6.2.2.2 Einordnung der Befunde

Für alle drei Mischverteilungsmodelle weisen die Schätzungen der Itemparameter große Ähnlichkeiten auf und für die Items am Testende unterscheiden sich die Schätzungen von denen aus einem Standard-IRT-Modell, das nicht für den Leistungsabfall kontrolliert. Dieser Befund passt zu den Ergebnissen einer Simulationsstudie von Suh et al. (2012), die das Modell von Bolt et al. (2002) mit dem Modell von Yamamoto (1995) hinsichtlich der Verbesserung der Itemparameterschätzung gegenüber einem Standard-IRT-Modell vergleichen, wenn ein Leistungsabfall in den Daten vorliegt. Insgesamt zeigt sich in der Studie von Suh et al. (2012), dass beide Modelle ähnlich gut geeignet sind, um die Verzerrung in den Itemschwierigkeiten durch Effekte des Leistungsabfalls zu verringern. Allerdings finden Suh et al. (2012), dass die Diskriminationsparameter im Modell von Yamamoto (1995) überschätzt wurden, während das Modell von Bolt et al. (2002) auch hier zu unverzerrten Schätzungen führte. In Studie 2 zeigten sich hingegen für die Itemdiskriminationen für alle drei Mischverteilungsmodelle sehr ähnliche Befunde.

Inwieweit das Modell von Bolt et al. (2002) geeignet ist, Leistungsabfall adäquat zu erfassen, wurde auch von Mittelhaëuser et al. (2013) diskutiert. Allerdings konzentriert sich deren Studie nur auf geringe Testbearbeitungsmotivation als Ursache für den Leis-

tungsabfall. Dazu vergleichen Mittelhaëuser et al. (2013) Selbstberichte zur Testbearbeitungsmotivation mit der Wahrscheinlichkeit, zur latenten Klasse mit Leistungsabfall zu gehören, und finden hier eine eher geringe Korrelation, was darauf hinweist, dass die identifizierten latenten Klassen im Modell von Bolt et al. (2002) nicht nur Unterschiede im Leistungsabfall widerspiegeln. Allerdings ist zu beachten, dass Mittelhaëuser et al. (2013) die Klassenwahrscheinlichkeiten nicht mit anderen Ursachen für Leistungsabfall wie etwa Ermüdung oder Zeitdruck in Beziehung gesetzt haben.

In dieser Studie zeigen Personen mit einer geringeren Mathematikfähigkeit einen größeren Leistungsabfall und an nicht-gymnasialen Schulformen zeigen mehr Schülerinnen und Schüler einen Leistungsabfall als am Gymnasium. Ähnliche Befunde finden sich in Studien zu Positionseffekten in Mathematiktests: Auch Debeer und Janssen (2013) berichten von einem negativen Zusammenhang von Mathematikfähigkeit und Leistungsabfall. Schulformunterschiede im Leistungsabfall zuungunsten der nicht-gymnasialen Schulformen finden auch G. Nagy, Lüdtke und Köller (2016). Ebenso wie in der vorliegenden Arbeit führt auch dort die Berücksichtigung des Leistungsabfalls im Modell zur Reduktion des Fähigkeitsunterschieds zwischen den Schulformen.

Zwei Simulationsstudien untersuchen, wie gut die Modelle von Bolt et al. (2002) und von Yamamoto (1995) den Bias durch Leistungsabfall reduzieren können, und vergleichen die Ergebnisse unter verschiedenen Kodierungen von Not-Reached-Items (Suh et al., 2006, 2012). Suh et al. (2006) finden, dass die Verzerrung in den Personenfähigkeiten geringer ist, wenn die Not-Reached-Items als fehlende Werte behandelt werden. In der Studie von Suh et al. (2012) zeigen sich für die Itemparameter keine Unterschiede in der Reduktion des Bias, wenn Not-Reached-Items als fehlende Werte oder als falsche Itemantworten behandelt werden. Zusammengefasst legen diese beiden Simulationsstudien nahe, dass sich in der vorliegenden Arbeit die Behandlung der Not-Reached-Items nicht auf die Schätzung der

Itemparameter auswirkt (s. Suh et al., 2012), aber es kann einen Effekt auf die Schätzung der Personenfähigkeiten geben (s. Suh et al., 2006).

6.2.3 Fazit

Ziel beider Studien war es, bestehende Ansätze für die Modellierung der Veränderung des Testbearbeitungsverhaltens zu erweitern und die Performanz in empirischen Datensätzen zu untersuchen. In beiden Studien wurden Daten aus Mathematiktests verwendet. Es zeigt sich, dass Testteilnehmende mit einer höheren Mathematikfähigkeit insgesamt tendenziell mehr Not-Reached-Items (also einen früheren Testabbruch) haben (Studie 1) aber einen geringeren Leistungsabfall zeigen (Studie 2). Die bestehenden Modelle wurden um kategorialen Kovariaten erweitert, um Gruppenunterschiede im Testbearbeitungsverhalten untersuchen zu können. Konsistent zeigen sich in beiden Studien differentielle Effekte zugunsten der jeweils leistungsstärkeren Gruppe (SuS am Gymnasium, Industriekaufleute) unter Kontrolle der Personenfähigkeiten. Wie Studie 2 zeigt, reduziert sich dadurch der Leistungsunterschied zwischen den Gruppen, wenn für den Leistungsabfall kontrolliert wird. Jedoch scheint in dieser Stichprobe die Reduktion im Hinblick auf den erheblichen Leistungsunterschied eher gering.

Zum Teil bestätigen, zum Teil widersprechen die Ergebnisse dieser Studien bisherigen Befunden, was insgesamt darauf deutet, dass die Veränderung des Testbearbeitungsverhaltens tiefer gehend untersucht werden sollte, etwa um bisher vernachlässigte Variablen zu identifizieren, die die Unterschiede (und Gemeinsamkeiten) in den Studienergebnissen erklären können.

In beiden Studien wird die Veränderung des Testbearbeitungsverhalten anhand anschaulicher Konzepte formalisiert und die untersuchten Modelle in einem gemeinsamen Rahmen verortet. In der ersten Studie werden die Modelle als Formen einer Survivalanalyse, die den Testabbruch untersucht, beschrieben. In der zweiten Studie wird ein generelles

Mischverteilungsmodell für den Leistungsabfall entwickelt. In beiden Studien ermöglicht der jeweilige Rahmen, die Unterschiede in den einzelnen Modellen auf einfache Weise darzustellen und zu kontrastieren, indem beschrieben wird, welche Restriktionen für welches Modell eingeführt werden.

Die Veränderung des Testbearbeitungsverhaltens wird in allen Modellen als abrupter Wechsel von der maximalen Leistung zu einer anderen Bearbeitungsstrategie definiert. Der Testabbruch stellt trivialerweise einen abrupten Wechsel dar, für den Leistungsabfall wird ebenfalls angenommen, dass er genau an einer Itemposition einsetzt. Andere Ansätze modellieren einen graduellen Leistungsabfall. Goegebeur et al. (2008) nehmen in ihrem Modell an, dass der Aufwand bei der Testbearbeitung zum Testende hin kontinuierlich abfällt. Testteilnehmende können sich sowohl im Startpunkt des Prozesses (d. h. in ihrem Schaltpunkt) als auch in der Stärke des Leistungsabfalls unterscheiden. Modelle für den Leistungsabfall mit einem abrupten Wechsel, wie sie in dieser Dissertation in der zweiten Studie verwendet werden, können auch als eine Approximation eines Modells mit graduellem Leistungsabfall angesehen werden (Jin & Wang, 2014). Der Vorteil ist dabei, dass diese Modelle weniger Parameter benötigen.

Die Vergleiche der verschiedenen Modelle für Not-Reached-Items oder Leistungsabfall untereinander und mit einem Standard-IRT-Modell deuten auf Unterschiede in Item- und Personenparametern und der Schätzung von Gruppenunterschieden in der Fähigkeit hin. Inwieweit diese Unterschiede auch praktisch bedeutsame Konsequenzen haben, muss abhängig von der Fragestellung bewertet werden.

Die Simulationsstudie in Studie 1 zeigt, dass der Effekt der Gruppenvariable auf die Personenfähigkeit sowie die Varianz der Fähigkeitsvariable unterschätzt wird, wenn differentielle Raten von Not-Reached-Items vorliegen, aber im Modell nicht berücksichtigt werden. Der Anteil der fehlenden Werte in den generierten Daten ist verglichen mit realen Datensätzen jedoch eher hoch, und es ist zu erwarten, dass die Verzerrung bei einer

geringeren Anzahl fehlender Werte kleiner ausfällt (vgl. Debeer et al., 2017; Glas & Pimentel, 2008) oder sogar unerheblich wird (vgl. Collins, Schafer & Kam, 2001; Pietsch, 2011). Durch den erheblichen Fähigkeitsunterschied zwischen Schulformen in Studie 2 wirkt sich die Berücksichtigung der Unterschiede im Leistungsabfall nur geringfügig auf den Gruppenunterschied aus.

Auch wenn die Auswirkungen auf die Schätzung von Fähigkeitsunterschieden in diesen Studien eher als gering zu bewerten sind, zeigen diese Analysen exemplarisch, wie diese Modelle im Rahmen von Sensitivitätsanalysen eingesetzt werden können, um abzuschätzen, wie sehr die Ergebnisse variieren, wenn bestimmte Annahmen zum Testbearbeitungsverhalten modelliert beziehungsweise ignoriert werden (vgl. Pohl & Carstensen, 2013).

Beide Studien zeigen darüber hinaus Anwendungen von Latent-Class-Modellen, die über deren ursprünglichen Zweck, die Kategorisierung von Personen, hinausgehen. Bauer (2005) unterscheidet bei der Anwendung von Mischverteilungsmodellen zwischen dem *direkten Ansatz* und dem *indirekten Ansatz*. Ersterer dient der Zuweisung von Personen zu einer der Kategorien der latenten Klassenvariable mit dem Ziel, die heterogene Gesamtpopulation zu entmischen und in homogene Teilgruppen zu zerlegen. Letzterer Ansatz versucht, die Heterogenität der Gesamtpopulation mithilfe der Klassenvariable zu beschreiben, ohne die Existenz qualitativ verschiedener Teilgruppen anzunehmen. Die Kategorisierung ist eher ein Werkzeug, um die unbekannte Verteilung der Variable in der Population beschreiben zu können.

Das MDSEM der ersten Studie weist Ähnlichkeiten mit dem indirekten Ansatz von Bauer (2005) auf. Die Kategorisierung in latente Klassen dient allein dem Zweck, die nicht-normalverteilte Missingness-Variable zu beschreiben. Die Mischverteilungsmodelle der zweiten Studie werden in dieser Arbeit ebenfalls eher im Sinne des indirekten Ansatzes verwendet, weil auch hier der Fokus auf der Beschreibung der Gesamtpopulation liegt. Jedoch wird das Modell von Bolt et al. (2002) auch dafür eingesetzt, um Testteilneh-

mende mit einem Leistungsabfall aus dem Datensatz zu filtern, um weitere Analysen nur auf Basis der Personen, die keinen Leistungsabfall zeigen, durchzuführen (z. B. Finn, 2015; Wollack et al., 2003). Dieses Vorgehen entspricht dem direkten Ansatz von Bauer (2005). Latent-Class-Modelle finden somit im Kontext der Modellierung des Testbearbeitungverhaltens vielfältigen Einsatz.

Die Modelle, die in dieser Dissertation vorgestellt werden, erlauben, das Ausmaß von Not-Reached-Items beziehungsweise Leistungsabfall vertiefend zu untersuchen. In den Studien wurde demonstriert, dass die Auswertung der Modelle keine Spezialsoftware benötigt, sondern mit Standardsoftware (hier Mplus; L. K. Muthén & Muthén, 1998-2012) durchgeführt werden kann, sodass die Modelle in einfacher Weise für verschiedene Fragestellungen eingesetzt werden können.

6.3 Anwendung der Modelle in Large-Scale-Assessments

Die vorgestellten Modelle dieser Dissertation sind für verschiedene Anwendungen im Kontext der Fähigkeitsmessung in LSAs relevant. Bei der Bewertung von Studienergebnissen erlaubt die Modellierung von Not-Reached-Items oder Leistungsabfall, die praktische Bedeutsamkeit einer Verzerrung in den Modellparametern, wenn die Veränderung des Testbearbeitungsverhaltens nicht berücksichtigt wird, zu untersuchen. Dies ist im Hinblick auf die Bewertung der Validität des Tests und der Testergebnisse relevant. Ein Bias in den Itemparametern kann zudem in Linkingstudien ein Problem darstellen (z. B. Mittelhaëuser et al., 2015a).

Auch zur Bewertung der Validität der Fähigkeitsunterschiede zwischen Gruppen ist die Veränderung des Testbearbeitungsverhaltens interessant, wenn sich zeigt, dass die Verteilung der Personenfähigkeit vom Ausmaß der Not-Reached-Items oder des Leistungsabfalls abhängt. Differentielle Raten von Not-Reached-Items und ein differentieller Leistungsab-

fall können dann zu gruppenspezifischen Verzerrungen der Fähigkeitsverteilung führen, was sich wiederum in einer Verschätzung der Fähigkeitsunterschiede der Gruppen zeigt. Darüber hinaus sprechen differentielle Effekte auch gegen die Fairness eines Tests, wenn das Testergebnis nicht nur von der Personenfähigkeit sondern auch von weiteren Personenmerkmalen abhängt (Camilli, 2006; Bridgeman, McBride & Monaghan, 2004; Evans & Reilly, 1972, 1973; Kok, 1988; Lu & Sireci, 2007; Tate, 2002).

Die Modelle ermöglichen, den Effekt von Not-Reached-Items und Leistungsabfall auf die Personenfähigkeit bei der Parameterschätzung zu berücksichtigen, um einen Bias in den Parametern zu reduzieren.[25] Die Mischverteilungsmodelle für den Leistungsabfall können außerdem dazu verwendet werden, um Personen mit einem Leistungsabfall zu identifizieren. Die Daten dieser Personen können dann für weitere Analysen aus dem Datensatz entfernt werden, um die Parameterschätzung nur auf Grundlage der Personen, die durchgängig maximale Leistung zeigen, durchzuführen (Finn, 2015; Wollack et al., 2003). Allerdings verändert sich dadurch die Fähigkeitsverteilung der Stichprobe, wenn Leistungsabfall und Personenfähigkeit zusammenhängen, sodass das Vorgehen nicht in allen Kontexten sinnvoll ist (vgl. S. L. Wise & DeMars, 2005).

Die Modelle können auch im Zuge der Testkonstruktion verwendet werden, etwa um die optimale Testlänge zu bestimmen, bevor ein substantieller Anteil der Testteilnehmenden den Test abbricht oder einen Leistungsabfall zeigt (vgl. Hartig & Buchholz, 2012). Durch die Analyse der Abbruchwahrscheinlichkeiten im Testverlauf kann das MDSEM dazu verwendet werden, um Items zu identifizieren, an denen die Wahrscheinlichkeit für einen Testabbruch besonders hoch ist. Anschließend kann analysiert werden, welche Eigenschaften der Items zu der hohen Abbruchrate führen, um diese Itemtypen gegebenenfalls von der finalen Testversion auszuschließen.

[25] Hier soll angemerkt sein, dass die Einschätzung, ob ein Bias vorliegt, theoretischer Begründungen bedarf. Vorausgesetzt wird dann, dass das Modell mit der Berücksichtigung der Veränderung des Testbearbeitungsverhaltens die Wahrheit der Testbearbeitung besser annähert als ein Standard-IRT-Modell.

Außerdem können die Modelle aber auch dazu genutzt werden, um die Veränderung des Testbearbeitungsverhaltens an sich näher zu untersuchen. Sowohl der Testabbruch als auch der Leistungsabfall stellen Personenmerkmale dar, die den Fokus einer Studie bilden können. Das individuelle Ausmaß von Not-Reached-Items und Leistungsabfall kann mit weiteren (kategorialen wie kontinuierlichen) Kovariaten in Verbindung gebracht werden (vgl. Cohen et al., 2002; Denis & Gilbert, 2012; Köhler et al., 2015a; Wild et al., 1982). Dies kann Aufschluss darüber geben, welche Prozesse der Veränderung des Testbearbeitungsverhaltens zugrunde liegen, und so helfen, ein besseres Verständnis von Testbearbeitungsstrategien zu erhalten.

6.4 Limitationen

Auch die vorliegende Arbeit weist eine Reihe von Limitationen auf, vor deren Hintergrund die Interpretation der Ergebnisse kritisch überprüft werden muss. Die Veränderung des Testbearbeitungsverhaltens wurde als Personenmerkmal betrachtet. Die Veränderung des Testbearbeitungsverhaltens ist auf Itemseite nur an die Itemeigenschaft *Position* gekoppelt (vgl. Debeer, 2016): Definitionsgemäß hängt es von der Position im Test ab, ob ein Item erreicht wird, und auch der Leistungsabfall wurde in dieser Arbeit so definiert, dass er an einer bestimmten Position im Test einsetzt.

Implizit wird in dieser Arbeit angenommen, dass keine anderen Itemmerkmale mit der Veränderung des Testbearbeitungsverhaltens zusammenhängen. Aber auch weitere Itemmerkmale können zu einer (individuell ausgeprägten) Veränderung des Testbearbeitungsverhaltens führen: Eine Reihe von Studien findet, dass Items, die als kognitiv aufwendiger wahrgenommen werden, und Item mit einem langen Text mit höherer Wahrscheinlichkeit geraten werden (DeMars, 2000; Goldhammer, Naumann & Greiff, 2015; Horst, 2010; Lee & Jia, 2014; S. L. Wise et al., 2009; Wolf et al., 1995). Daneben zeigt sich, dass Items

mit einem offenen Antwortformat (Koretz et al., 1993) oder mit höheren Schwierigkeiten (Debeer et al., 2017; Koretz et al., 1993) häufiger ausgelassen werden. Bei entsprechender Anordnung der Items im Test können sich Leistungsabfall oder Testabbruch und Effekte der Itemmerkmale überlagern und gegenseitig begünstigen, wenn etwa die Items nach Format oder Schwierigkeit sortiert sind (vgl. Lawrence, 1993; Koretz et al., 1993). Wenn es sich nicht um ein Multi-Matrix-Design handelt, bei dem die Lösungswahrscheinlichkeiten eines Items zu unterschiedlichen Positionen verglichen werden können, lassen sich diese Effekte nicht trennen (vgl. Debeer, 2016).

Eine grundlegende Annahme aller in dieser Dissertation verwendeten Modelle ist, dass die Testteilnehmenden die Items in der vorgegebenen Reihenfolge bearbeiten. Zwar wird diese Annahme durch viele Befunde zum monotonen Leistungsabfall (z. B. Hartig & Buchholz, 2012) und vermehrten Rateverhalten am Testende (z. B. Schnipke & Scrams, 1997; S. L. Wise, 2006) gestützt, aber es gibt auch Hinweise darauf, dass schwierige Items eher geraten (Lee & Jia, 2014) oder ausgelassen werden (Debeer et al., 2017), was darauf hindeuten kann, dass Testteilnehemende zuerst die als einfach wahrgenommenen Items bearbeiten, sodass sich ein Leistungsabfall vor allem bei den schwierigen Items zeigt (Bejar, 1985; Y.-W. Chang et al., 2014; J. Chang et al., 2016; Cao & Stokes, 2008) oder dass Personen die schwierigen Items nicht mehr bearbeiten und stattdessen auslassen. Darüber hinaus finden Stenlund, Eklöf und Lyrén (2016), dass Testteilnehmende mit geringerer Personenfähigkeit eher den Test in der vorgegebenen Reihenfolge bearbeiten, während Testteilehmende mit höherer Fähigkeit die Bearbeitungsreihenfolge frei bestimmen.

Die Modelle basieren außerdem auf der Annahme, dass Testteilnehmende zu Beginn des Tests mit maximaler Leistung arbeiten. Es ist aber gerade in Low-Stakes-Tests denkbar, dass Personen bereits zu Beginn des Tests keine maximale Leistung zeigen (s. S. L. Wise & DeMars, 2005; S. L. Wise & Kong, 2005). Ursache dafür kann eine geringe Testbearbeitungsmotivation sein. Auch Ermüdung kann eine Ursache darstellen, zum Beispiel wenn

vor dem untersuchten Test bereits mehrere andere Tests durchgeführt wurden (Barry & Finney, 2016; DeMars & Wise, 2007).[26]

Die Anwendung von Mischverteilungsmodellen zur Erfassung des Leistungsabfalls, setzt aber voraus, dass Testteilnehmende zumindest auf dem ersten Item maximale Leistung zeigen. Für Personen, die bereits am Testanfang mit geringerer Leistung arbeiten, kann unter Umständen kein Leistungsabfall identifiziert werden, sodass sie der Gruppe von Personen ohne Leistungsabfall zugeordnet werden (Cao & Stokes, 2008; Cohen et al., 2002; Goegebeur et al., 2008; Wollack, Suh & Bolt, 2007). Damit enthält die latente Klasse aber nicht mehr, wie in den Modellen angenommen, nur die Itemantworten unter maximaler Leistung, sodass die Klassenzuordnung nicht allein die Variation im Leistungsabfall darstellt. Aussagen über den Zusammenhang von Personenfähigkeit und Leistungsabfall werden dadurch verzerrt (Cao & Stokes, 2008). Auch für Personen mit geringer Fähigkeit kann die Klassenzuordnung ungenau sein, weil die Identifikation eines Leistungsabfalls einen hinreichend deutlichen Unterschied in den Itemantworten vor und nach dem Schaltpunkt voraussetzt, der bei Testteilnehmende mit sehr geringer Personenfähigkeit nicht gegeben ist, wenn sie bereits zu Beginn des Tests eine geringe Leistung zeigen (Cao & Stokes, 2008; Cohen et al., 2002; Yamamoto & Everson, 1997). Auch dieser Umstand kann dazu führen, dass latenten Klassen nicht nur den Grad von Leistungsabfall abbilden, sodass die Verteilung des Leistungsabfalls verschätzt wird.

Ein generelles Problem von Mischverteilungsmodellen ist, dass die Identifikation von latenten Klassen sensitiv gegenüber Fehlspezifikationen des zugrunde liegenden IRT-Modells ist. Alexeev, Templin und Cohen (2011) zeigen in einer Simulationsstudie, dass ein Mischverteilungsmodell mehrere latente Klassen implizieren kann, wenn das zugrunde liegende IRT-Modell zu einfach ist, etwa wenn ein 1PL-Mischverteilungsmodell auf Daten angepasst

[26] Es ist ebenso denkbar, dass Personen aus diesen Gründen den Test im Ganzen nicht bearbeiten (*unit nonresponse*, vgl. Enders, 2010; Schafer & Graham, 2002). Das stellt allerdings ein anderes *Missing-Data*-Problem dar, das nicht im Fokus dieser Dissertation steht.

wird, die einem 2PL oder 3PL entsprechen. Die identifizierten latenten Klassen modellieren dann keine wahre Parameterheterogenität in den Daten, sondern vielmehr wird über die Klassen die unberücksichtigte Modellkomplexität ausgeglichen (Alexeev et al., 2011). Wie Y.-F. Chen und Jiao (2013) in ihrer Simulationsstudie zeigen, lässt sich der Effekt der fälschlicherweise identifizierten latenten Klassen im 1PL-Mischverteilungsmodell vor allem auf eine Unterschätzung der Itemdiskriminationsparameter zurückführen.

Doch auch wenn die untersuchte Stichprobe tatsächlich Heterogenitäten aufweist, kann eine Fehlspezifikation des IRT-Modells zu einer ungenauen oder möglicherweise falschen Klassenlösung führen. Cho, Cohen und Kim (2014) haben mit einem 1PL-Mischverteilungsmodell Leistungsabfall in einem Test modelliert und vergleichen, wie die Interpretation der latenten Klassen variiert, wenn statt eines eindimensionalen 1PL ein Testletmodell zugrunde gelegt wird, das durch das Testdesign bedingte lokale Abhängigkeiten zwischen den Items berücksichtigt (vgl. Bradlow, Wainer & Wang, 1999; Wainer, Bradlow & Wang, 2007). Cho et al. (2014) legen für den Leistungsabfall das *Mixture-Rasch-Modell* von Bolt et al. (2002) zugrunde (s. Kap. 2.4, S. 31ff.). Im 1PL-Mischverteilungsmodell wird der Anteil der Personen mit Leistungsabfall höher geschätzt, als wenn das Testlet-Mischverteilungsmodell verwendet wird. Daneben führt das Testlet-Mischverteilungsmodell zu einer plausibleren Interpretation für den Leistungsabfall.

Die Studien dieser Dissertation beschäftigen sich mit je einem Aspekt der Veränderung des Testbearbeitungsverhaltens. In LSAs treten normalerweise Not-Reached-Items und Leistungsabfall gleichzeitig auf. Studien finden, dass beide Phänomene positiv miteinander korreliert sind (De Boeck et al., 2011; Schnipke & Scrams, 2002; Wang, 2011), und in einigen Ansätze wird vorgeschlagen, beide Phänomene zu erfassen (Cho et al., 2014; Mroch & Bolt, 2006; L. L. Wise, 1996). In der ersten Studie wird die Fähigkeit über alle bearbeiteten Items geschätzt und damit wird implizit angenommen, dass kein Leistungsabfall vorliegt. Die Fähigkeitsschätzung kann durch den unberücksichtigten Leistungsabfall verzerrt

sein, was ebenfalls zu Verzerrungen in der Schätzung des Zusammenhangs mit Testabbruch und den differentiellen Effekten der Gruppenzugehörigkeit geführt haben kann.

Im Datensatz, der in der zweiten Studie verwendet wurde, sind die Not-Reached-Items als falsche Antwort kodiert, sodass der identifizierte Leistungsabfall sowohl auf eine wahre Reduktion in der Lösungswahrscheinlichkeit als auch auf Not-Reached-Items zurückgeführt werden kann, aber beide Aspekte nicht getrennt werden können. Das Ausmaß des Leistungsabfalls kann daher auch auf Not-Reached-Items zurückzuführen sein. Ferner ist es möglich, dass sich beide Gruppen substantiell in der Ausprägung beider Phänomene, Not-Reached-Items und Leistungsabfall, unterscheiden, sodass in der einen Gruppe Not-Reached-Items und in der anderen Gruppe Leistungsabfall überwiegen. Da sich allerdings die Gruppen nur in der Anzahl der betroffenen Personen aber nicht im Ausmaß des Leistungsabfalls unterscheiden, erscheint es im Datenbeispiel der zweiten Studie unwahrscheinlich, dass es in der einen Gruppe Not-Reached-Items und in der anderen ein tatsächlicher Leistungsabfall deutlich überwiegen.

Zusammenfassend lässt sich sagen, dass die Interpretation der Ergebnisse davon abhängt, wie gut die Modelle in der Lage sind, die Veränderung des Testbearbeitungsverhaltens und auch die Personenfähigkeit adäquat zu erfassen (vgl. Keizer-Mittelhaëuser, 2014, Kap. 6). Fehlspezifikationen können zu Fehlinterpretationen führen. Wenn möglich, sollten die gewählten Modelle mit komplexeren Modellen verglichen werden, um die Sensitivität der Parameter zu überprüfen (vgl. Cho et al., 2014).

6.5 Zukünftige Forschungsfragen

Wie in Abschnitt 6.3 erläutert, bieten die in dieser Dissertation vorgeschlagenen Modelle verschiedene Anwendungsmöglichkeiten in LSAs. Die Studien dieser Dissertation liefern erste Ergebnisse zu Korrelaten von Leistungsabfall und Not-Reached-Items und zu de-

ren Zusammenhängen mit der Personenfähigkeit. Die Untersuchung von Korrelaten kann Aufschluss über die Ursachen der Veränderung des Testbearbeitungsverhaltens geben. Beispielsweise kann mit dem Einbezug motivationaler Variablen die Hypothese überprüft werden, dass Leistungsabfall und Not-Reached-Items Ausdruck geringer Testbearbeitungsmotivation sind (vgl. DeMars, 2002; S. L. Wise & DeMars, 2005). Daneben können differentielle Effekte für weitere Gruppierungsmerkmale untersucht werden. Wie die Zusammenschau der Befunde der ersten Studie und der Ergebnissen von Köhler et al. (2015a) zeigen, können die Zusammenhänge in Abhängigkeit des Alters der Testteilnehmenden variieren, was in weiteren Studien genauer untersucht werden könnte.

Studien zeigen, dass zwischen verschiedenen Testdomänen das Ausmaß von Leistungsabfall variiert (Debeer & Janssen, 2013; G. Nagy, Lüdtke & Köller, 2016) und sich der Zusammehang von Not-Reached-Items und Personenfähigkeit unterscheidet (Glas & Pimentel, 2008; Köhler et al., 2015a; Pohl et al., 2014). Die Effekte sind dabei in Tests zum Leseverständnis stärker als in Tests zur Mathematikfähigkeit ausgesprägt (Debeer & Janssen, 2013; G. Nagy, Lüdtke & Köller, 2016; Pohl et al., 2014). Vor diesem Hintergrund ist zu erwarten, dass die in dieser Dissertation vorgestellten Modelle in Daten zum Leseverständnis größere Effekte zeigen, was in nachfolgenden Studien untersucht werden könnte. Ferner sollten in Simulationsstudien die Auswirkungen von Fehlspezifikationen des Messmodells für die Fähigkeit weiter untersucht werden (vgl. Cho et al., 2014), um zu genaueren Aussagen zu gelangen, unter welchen Bedingungen mit Verzerrungen zu rechnen ist.

6.6 Resümee

Ein häufig verwendetes Zitat in Arbeiten in der pädagogisch-psychologischen Methodenforschung stammt von Box und Draper (1987): „*[Essentially,] all models are wrong, but some are useful.*“(Box & Draper, 1987, S. 424). Auch im Kontext der Modellierung von Leistungs-

abfall und Testabbruch und deren Bedeutung für die Parameterschätzung in Leistungstests hat dieses Zitat eine (gewisse) Gültigkeit. Letztendlich kann der Effekt der Veränderung des Testbearbeitungsverhaltens nicht isoliert als einzige Quelle von Störvarianz im Messmodell betrachtet werden. Anliegen dieser Dissertation ist vielmehr, die Bedeutung von Annahmen über das Testbearbeitungsverhalten zu betonen und mit den vorgestellten Ansätzen Modelle vorzuschlagen, die ermöglichen, die Phänomene Not-Reached-Items und Leistungsabfall zu erfassen, und die zugleich einen definitorischen Rahmen abstecken, innerhalb dessen die Veränderung des Testbearbeitungsverhaltens konzeptualisiert werden kann.

Wie alle Modelle sind auch diese Modelle aber zu wenig komplex, um die Wirklichkeit und die Bandbreite des Testbearbeitungsverhaltens ganz zu erfassen (vgl. Atkinson & Gray, 2006, 94f.). Aber sie bieten einen Ansatzpunkt, um zu untersuchen, inwieweit die Studienergebnisse robust sind, wenn statt eines Standard-IRT-Modells erweiterte und (vermutlich) plausiblere Annahmen zum Testbearbeitungsverhalten in das Modell integriert werden. So können die Modelle, die in dieser Dissertation vorgestellt werden, dazu verwendet werden, um Erkenntnisse über die Bedeutsamkeit einer Fehlspezifikation des Standard-IRT-Modells zu gewinnen (vgl. Sinharay & Haberman, 2014). Der Einsatz der Modelle sollte dabei von den Erwartungen und Theorien zum Testbearbeitungsverhalten in der jeweiligen Testsituation geleitet und kritisch mit anderen Operationalisierungen verglichen werden.

Literaturverzeichnis

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*, 117-128. doi: 10.1111/j.0006-341X.1999.00117.x

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317-332. doi: 10.1007/BF02294359

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

Albert, P. S. & Follmann, D. A. (2009). Shared-parameter models. In G. Fitzmaurice, M. Davidian, G. Verbeke & G. Molenberghs (Hrsg.), *Longitudinal data analysis* (S. 433-452). Boca Raton, FL: Chapman & Hall/CRC.

Alexeev, N., Templin, J. & Cohen, A. S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*, *48*, 313-332. doi: 10.1111/j.1745-3984.2011.00146.x

Allemann-Ghionda, C. (2003). Klasse, Gender oder Ethnie? Zum Bildungserfolg von Schüler/innen mit Migrationshintergrund. Von der Defizitperspektive zur Ressourcenorientierung. *Zeitschrift für Pädagogik*, *52*, 350-362.

Allison, P. D. (2014). *Event history and survival analysis* (2. Aufl.). Los Angeles, CA: Sage.

Allmendinger, J., Ebner, C. & Nikolai, R. (2010). Soziologische Bildungsforschung. In

R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (3. Aufl., S. 47-70). Wiesbaden: VS Verlag für Sozialwissenschaften.

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*, 5-14.

Asseburg, R. (2011). *Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests* (Unveröffentlichte Dissertation). Christian-Albrechts-Universität zu Kiel.

Atkinson, Q. D. & Gray, R. D. (2006). How old is the indo-european language family? Illumination or more moths to the flame? In P. Forster & C. Renfrew (Hrsg.), *Phylogenetic methods and the prehistory of languages* (Kap. 8). Cambridge, UK: McDonald Institute for Archaeological Research.

Bacci, S. & Bartolucci, F. (2015). A multidimensional finite mixture structural equation model for nonignorable missing responses to test items. *Structural Equation Modeling*, *22*, 352-365. doi: 10.1080/10705511.2014.937376

Barry, C. L. & Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurment in Education*, *29*, 46-64. doi: 10.1080/08957347.2015.1102914

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R. & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*, 342-363. doi: 10.1080/15305058.2010.508569

Bartholomew, D., Knott, M. & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3. Aufl.). Chichester, UK: Wiley.

Barton, M. A. & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Bericht Nr. RR-81-20). Princeton, NJ: Educational Testing Service.

Bauer, D. J. (2005). A semiparametric approach to modeling nonlinear relations among latent variables. *Structural Equation Modeling*, *12*, 513-535. doi: 10.1207/s15328007sem1204_1

Baumert, J. (2016). Leistungen, Leistungsfähigkeit und Leistungsgrenzen der empirischen Bildungsforschung: Das Beispiel von Large-Scale-Assessment-Studien zwischen Wissenschaft und Politik. *Zeitschrift für Erziehungswissenschaft*, *19*, 215-253. doi: 10.1007/s11618-016-0704-4

Baumert, J. & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*, 441-462. doi: 10.1007/BF03173192

Baumert, J., Maaz, K. & Trautwein, U. (2009). Editorial – Sonderheft 12: Bildungsentscheidungen. *Zeitschrift für Erziehungswissenschaft*, *12*, 7-10.

Becker, M., Lüdtke, O., Trautwein, U. & Baumert, J. (2006). Leistungszuwachs in Mathematik: Evidenz für einen Schereneffekt im mehrgliedrigen Schulsystem? *Zeitschrift für Pädagogische Psychologie*, *20*, 233-242. doi: 10.1024/1010-0652.20.4.233

Bejar, I. I. (1985). *Test speededness under number right scoring: An analysis of the Test Of English as a Foreign Language* (Bericht Nr. RR-85-11). Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Hrsg.), *Statistical theories of mental test scores* (S. 397-549). Reading, MA: Addison-Wesley.

Bolt, D. M., Cohen, A. S. & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331-348. doi: 10.1111/j.1745-3984.2002.tb01146.x

Bolt, D. M., Mroch, A. A. & Kim, J.-S. (2003). *An empirical investigation of the hybrid*

IRT model for improving item parameter estimation in speeded tests. Paper Presented at the Annual meeting of the American Educational Research Association (AERA). Chicago, IL.

Boughton, K. A. & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier & C. H. Carstensen (Hrsg.), *Multivariate and mixture distribution rasch models: extensions and applications* (S. 147-156). New York, NY: Springer.

Box, G. E. P. & Draper, N. R. (1987). *Empirical model-building and response surfaces.* New York, NY: John Wiley & Sons.

Bradlow, E. T. & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics*, *23*, 236-243. doi: 10.3102/10769986023003236

Bradlow, E. T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168. doi: 10.1007/BF02294533

Bridgeman, B. & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement*, *41*, 137-148. doi: 10.1111/j.1745-3984.2004.tb01111.x

Bridgeman, B., McBride, A. & Monaghan, W. (2004). *Testing and time limits* (R&D Connections). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Trapani, C. & Curley, E. (2004). Impact of fewer questions per section on SAT I Scores. *Journal of Educational Measurement*, *41*, 291-310. doi: 10.1111/j.1745-3984.2004.tb01167.x

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*, 1-20. doi: 10.1186/s40536-017-0042-x

Camilli, G. (2006). Test fairness. In R. Brennan (Hrsg.), *Educational measurement* (4. Aufl., S. 221-256). Westport, CT: Praeger.

Cao, J. & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*, 209-230. doi: 10.1007/s11336-007-9045-9

Chang, J., Tsai, H., Su, Y.-H. & Lin, E. M. H. (2016). A three-parameter speeded item response model: Estimation and application. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas & M. Wiberg (Hrsg.), *Quantitative psychology research: The 80th annual meeting of the psychometric society* (S. 27-38). New York, NY: Springer.

Chang, Y.-W., Tsai, R.-C. & Hsu, N.-J. (2014). A speeded item response model: Leave the harder till later. *Psychometrika*, *79*, 255-274. doi: 10.1007/s11336-013-9336-2

Chen, H. H., von Davier, M., Yamamoto, K. & Kong, N. (2015). *Comparing data treatments on item-level nonresponse and their effects on data analysis of large-scale assessments: 2009 PISA study* (Bericht Nr. RR-15-12). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12059

Chen, Y.-F. & Jiao, H. (2013). Does model misspecification lead to spurious latent classes? An evaluation of model comparison indices. In R. E. Millsap, L. A. van der Ark, D. M. Bolt & C. M. Woods (Hrsg.), *New developments in quantitative psychology* (S. 345-355). New York, NY: Springer. doi: 0.1007/978-1-4614-9348-8__23

Cho, S.-J., Cohen, A. S. & Kim, S.-H. (2014). A mixture group bifactor model for binary responses. *Structural Equation Modeling*, *21*, 375-395. doi: 10.1080/10705511.2014.915371

Cohen, A. S., Wollack, J. A., Bolt, D. M. & Mroch, A. A. (2002). *A mixture Rasch model analysis of test speededness.* Paper presented at the annual meeting of the American Educational Research Association (AERA). New Orleans, LA.

Cokley, K. (2007). Critical issues in the measurement of ethnic and racial identity: A referendum on the state of the field. *Journal of Counseling Psychology*, *54*, 224-234. doi: 10.1037/0022-0167.54.3.224

Collins, L. M. & Lanza, S. T. (2010). *Latent class and latent transition analysis: With*

applications in the social, behavioral, and health sciences. Hoboken, NJ: Wiley.

Collins, L. M., Schafer, J. L. & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330-351. doi: 10.1037//1082-989X.6.4.330

Culpepper, S. A. (2017). The prevalence and implications of slipping on low-stakes, large-scale assessments. *Journal of Educational and Behavioral Statistics*, *42*, 706-725. doi: 10.3102/1076998617705653

Davey, T. & Lee, Y.-H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised general test* (Bericht Nr. GREB-08-01). Princeton, NJ: Educational Testing Service.

Debeer, D. (2016). *Item-position effects and missing responses in large-scale assessments: Models and applications* (Unveröffentlichte Dissertation). Katholieke Universiteit Leuven, B.

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. doi: 10.3102/1076998614558485

Debeer, D. & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*, 164-185. doi: 10.1111/jedm.12009

Debeer, D., Janssen, R. & De Boeck, P. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement*, *54*, 333-363. doi: 10.1111/jedm.12147

De Boeck, P., Cho, S.-J. & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement*, *35*, 583-603. doi: 10.1177/0146621611428446

De Boeck, P. & Wilson, M. (Hrsg.). (2004). *Explanatory item response models: A genera-*

lized linear and nonlinear approach. New York, NY: Springer.

Dedering, K. & Holtappels, H. G. (2010). Schulische Bildung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (3. Aufl., S. 365-382). Wiesbaden: VS Verlag für Sozialwissenschaften.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*, 55-77. doi: 10.1207/s15324818ame1301_3

DeMars, C. E. (2002). *Missing data and IRT item parameter estimation.* Paper presented at the annual meeting of the American Educational Research Association (AERA). Chicago, IL.

DeMars, C. E., Bashkov, B. M. & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Reserach & Practice in Assessment*, *8*, 69-82.

DeMars, C. E. & Wise, S. L. (2007). *Can differential rapid-guessing behavior lead to differential item functioning?* Paper presented at the annual meeting of the American Educational Research Association (AERA). Chicago, IL.

Denis, P. L. & Gilbert, F. (2012). The effect of time constraints and personality facets on general cognitive ability (GCA) assessment. *Personality and Individual Differences*, *52*, 541-545. doi: 10.1016/j.paid.2011.11.024

Diefenbach, H. (2007). Schulerfolg von ausländischen Kindern und Kindern mit Mitgrationshintergrund als Ergebnis individueller und institutioneller Faktoren. In *Migrationshintergrund von Kindern und Jugendlichen: Wege zur Weiterentwicklung der amtlichen Statistik* (S. 43-54). Bonn: Bundesministerium für Bildung und Forschung.

Diggle, P. & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *43*, 49-93. doi: 10.2307/2986113

Dorans, N. J., Pommerich, M. & Holland, P. W. (Hrsg.). (2007). *Linking and aligning*

scores and scales. New York, NY: Springer.

Dorans, N. J., Schmitt, A. P. & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, *29*, 309-319. doi: 10.1111/j.1745-3984.1992.tb00379.x

Educational Testing Service. (o. J.). *The TOEFL®test.* Princeton, NJ. Zugriff auf `https://www.ets.org/toefl/` (Zugriff am 16.09.2017)

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*, 311-326. doi: 10.1080/15305050701438074

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*, 345-356. doi: 10.1080/0969594X.2010.516569

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

Emenogu, B. C. & Childs, R. A. (2005). Curriculum, translation, and differential functioning of measurement and geometry items. *Canadian Journal of Education*, *28*, 128-146.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.

Evans, F. R. & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement*, *9*, 123-131. doi: 10.1111/j.1745-3984.1972.tb00767.x

Evans, F. R. & Reilly, R. R. (1973). A study of test speededness as a potential source of bias in the quantitative score of the admission test for graduate study in business. *Research In Higher Education*, *1*, 173-183. doi: 10.1007/BF00991339

Finch, W. H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*, 225-245.

Finch, W. H. & French, B. F. (2012). Parameter estimation with mixture item response theory models: A monte carlo comparison of maximum likelihood and Bayesi-

an methods. *Journal of Modern Applied Statistical Methods*, *11*, 167-178. doi: 10.22237/jmasm/1335845580

Finn, B. (2015). *Measuring motivation in low-stakes assessments* (Bericht Nr. RR-15-19). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12067

Formann, A. K. (1984). *Die Latent-Class-Analyse: Einführung in Theorie und Anwendung.* Weinheim: Beltz.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology*, *38*, 87-111. doi: 10.1111/j.2044-8317.1985.tb00818.x

Foy, P., Brossman, B. & Galia, J. (2012). Scaling the TIMSS and PIRLS 2011 achievement data. In M. O. Martin & I. V. S. Mullis (Hrsg.), *Methods and procedures in TIMSS and PIRLS 2011.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Foy, P. & Yin, L. (2017). Scaling the PIRLS 2016 achievement data. In M. O. Martin, I. V. Mullis & M. Hooper (Hrsg.), *Methods and procedures in PIRLS 2016* (S. 12.1-12.38). Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

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

Geier, T. & Zaborowski, K. U. (Hrsg.). (2016). *Migration: Auflösungen und Grenzziehungen – Perspektiven einer erziehungswissenschaftlichen Migrationsforschung.* Wiesbaden: Springer VS.

Geiser, C., Lehmann, W. & Eid, M. (2006). Separating "rotators" from "nonrotators" in the mental rotations test: A multigroup latent class analysis. *Multivariate Behavioral Research*, *41*, 261-293. doi: 10.1207/s15327906mbr4103_2

Ghahramani, Z., Griffiths, T. L. & Sollich, P. (2006). Bayesian nonparametric latent feature models. In J. M. Bernardo et al. (Hrsg.), *Proceedings of the eighth Valencia*

international meeting. Oxford, UK: Oxford University Press.

Glas, C. A. W. & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*, 907-922. doi: 10.1177/0013164408315262

Glas, C. A. W., Pimentel, J. L. & Lamers, S. M. A. (2015). Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling*, *57*, 523-541.

Goegebeur, Y., De Boeck, P., Wollack, J. A. & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*, 65-87. doi: 10.1007/S11336-007-9031-2

Goldhammer, F., Martens, T., Christoph, G. & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers Nr. 133). Paris, F: OECD. doi: 10.1787/5jlzfl6fhxs2-en

Goldhammer, F., Naumann, J. & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence*, *3*, 21-40. doi: 10.3390/jintelligence3010021

Gonzalez, E. & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments [IEA-ETS Research Institute Monograph]. In M. von Davier & D. Hastedt (Hrsg.), *IERI monograph series: Issues and methodologies in large-scale assessments* (Bd. 3, S. 125-156). International Association for the Evaluation of Educational Achievement (IEA) and Educational Testing Service (ETS).

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576. doi: 10.1146/annurev.psych.58.110405.085530

Guill, K., Lüdtke, O. & Köller, O. (2017). Academic tracking is related to gains in students' intelligence over four years: Evidence from a propensity score matching

study. *Learning and Instruction*, *47*, 43-52. doi: 10.1016/j.learninstruc.2016.10.001

Guo, J., Wall, M. & Amemiya, Y. (2006). Latent class regression on latent factors. *Biostatistics*, *7*, 145-163. doi: 10.1093/biostatistics/kxi046

Hailey, E., Callahan, C. M., Azano, A. & Moon, T. R. (2012). An evaluation of test speededness in an assessment for third-grade gifted students. *Journal of Advanced Academics*, *23*, 292-304. doi: 10.1177/1932202X12462575

Haladyna, T. M. & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*, 17-27. doi: 10.1111/j.1745-3992.2004.tb00149.x

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer – Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Harel, O. & Schafer, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika*, *96*, 37-50. doi: 10.1093/biomet/asn069

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.

Hecht, M. (2015). *Optimierung von Messinstrumenten im Large-scale Assessment* (Unveröffentlichte Dissertation). Humboldt Universität zu Berlin.

Hecht, M., Weirich, S., Siegle, T. & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement*, 1-24. doi: 10.1177/0013164415573311

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.

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

Holman, R. & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1-17. doi: 10.1348/000711005X47168

Horst, S. J. (2010). *A mixture-modeling approach to exploring test-taking motivation in large-scale low-stakes contexts* (Unveröffentlichte Dissertation). James Madison University, Harrisonburg, VA.

Hutchison, D. & Yeshanew, T. (2009). Augmenting the use of the Rasch model under time constraints. *Quality & Quantity*, *43*, 717-729. doi: 10.1007/s11135-007-9156-5

Jin, K.-Y. & Wang, W.-C. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement*, *51*, 178-200. doi: 10.1111/jedm.12041

Kao, G. & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. *Annual Review of Sociology*, *29*, 417-442. doi: 10.1146/annurev.soc.29.010202.100019

Kato, K. (2016). Measurement issues in large-scale educational assessments. *The Annual Report of Educational Psychology in Japan*, *55*, 148-164. doi: 10.5926/arepj.55.148

Keizer-Mittelhaëuser, M.-A. (2014). *Modeling the effect of differential motivation on linking educational tests* (Unveröffentlichte Dissertation). Tilburg University, NL.

Kim, M., Vermunt, J., Bakk, Z., Jaki, T. & Horn, M. L. V. (2016). Modeling predictors of latent classes in regression mixture models. *Structural Equation Modeling*, *23*, 601-614. doi: 10.1080/10705511.2016.1158655

Köhler, C. (2015). *Isn't something missing? Latent variable models accounting for item nonresponse* (Unveröffentlichte Dissertation). Freie Universität Berlin.

Köhler, C., Pohl, S. & Carstensen, C. H. (2015a). Investigating mechanisms for missing

responses in competence tests. *Psychological Test and Assessment Modeling*, *57*, 499-522. doi: 10.1177/0013164414561785

Köhler, C., Pohl, S. & Carstensen, C. H. (2015b). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement*, *75*, 850-874. doi: 10.1177/0013164414561785

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Hrsg.), *Latent trait and latent class models* (S. 263-275). New York, NY: Plenum Press.

Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, *28*, 219-226. doi: 10.1177/0146621604265030

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

Kolen, M. J., Tong, Y. & Brennan, R. L. (2011). Scoring and scaling educational tests. In A. A. von Davier (Hrsg.), *Statistical models for test equating, scaling, and linking* (S. 43-58). New York, NY: Springer. doi: 10.1007/978-0-387-98138-3

Köller, O. (1998). *Zielorientierungen und schulisches Lernen.* Münster: Waxmann.

Koretz, D., Lewis, E., Skewes-Cox, T. & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress* (CSE Technical Report Nr. 357). Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.

Kuha, J., Katsikatsou, M. & Moustaki, I. (2018). Latent variable modelling with non-ignorable item non-response: Multigroup response propensity models for cross-national analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *online first*. doi: 10.1111/rssa.12350

Kultusministerkonferenz. (2016). *Gesamtstrategie der Kultusministerkonferenz zum Bil-*

dungsmonitoring. Berlin.

Langeheine, R. & Rost, J. (Hrsg.). (1988). *Latent trait and latent class models.* New York, NY: Plenum Press.

Law School Admission Council. (o. J.). *Law School Admissions Test (LSAT).* Newtown, PA. Zugriff auf `https://www.lsac.org/jd/lsat/about-the-lsat` (Zugriff am 16.09.2017)

Lawrence, I. M. (1993). *The effect of test speededness on subgroup performance* (Bericht Nr. RR-93-49). Princeton, NJ: Educational Testing Service.

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

Lee, Y.-H. & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, *2*, 1-24. doi: 10.1186/s40536-014-0008-1

Lehmann, R. H. & Peek, R. (2011). LAU 5: Aspekte der Lernausgangslage und der Lernentwicklung – Klassenstufe 5: Ergebnisse einer längsschnittlichen Untersuchung in Hamburg im September 1996. In Behörde für Schule und Berufsbildung (Hrsg.), *LAU – Aspekte der Lernausgangslage und der Lernentwicklung: Klassenstufen 5, 7 und 9.* Münster: Waxmann.

Li, F., Cohen, A. S., Kim, S.-H. & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*, 353-373. doi: 10.1177/0146621608326422

Lilley, M., Barker, T. & Britton, C. (2004). The development and evaluation of a software prototype for computer-adaptive testing. *Computers & Education*, *43*, 109-123. doi: 10.1016/j.compedu.2003.12.008

Lindner, C., Nagy, G., Ramos Arhuis, W. A. & Retelsdorf, J. (2017). A new perspective on

the interplay between self-control and cognitive performance: Modeling progressive depletion patterns. *PLoS ONE*, *12*, 1-22. doi: 10.1371/journal.pone.0180149

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, *81*, 471-483. doi: 10.1093/biomet/81.3.471

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112-1121. doi: 10.1080/01621459.1995.10476615

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data* (2. Aufl.). New York, NY: Wiley.

Livingston, S. A. (2014). *Equating test scores (without IRT)* (2. Aufl.). Princeton, NJ. (no doi)

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*, 247-264. doi: 10.1007/BF02291471

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

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, *48*, 477-482. doi: 10.1007/BF02293689

Lu, Y. & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *Winter*, 29-37. doi: 10.1111/j.1745-3992.2007.00106.x

Lubke, G. H. & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*, 21-39. doi: 10.1037/1082-989X.10.1.21

Ludlow, L. H. & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement*, *59*, 615-630. doi: 10.1177/0013164499594004

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung: Probleme und Lösungen. *Psychologische*

Rundschau, *58*, 103-117. doi: 10.1026/0033-3042.58.2.103

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, *37*, 304-315. doi: 10.1177/0146621613475471

Martin, M. O., Mullis, I. V. & Hooper, M. (Hrsg.). (2016). *Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Masyn, K. E. (2009). Discrete-time survival factor mixture analysis for low-frequency recurrent event histories. *Research in Human Development*, *6*, 165-194. doi: 10.1080/15427600902911270

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.

Michaelides, M. P. (2010). A review of the effects on IRT item parameter estimates with a focus on misbehaving common items in test equating. *Frontiers in Psychology*, *1*, 1-7. doi: 10.3389/fpsyg.2010.00167

Mislevy, R. J., Johnson, E. G. & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics*, *17*, 131-154. doi: 10.3102/10769986017002131

Mislevy, R. J. & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195-245. doi: 10.1007/BF02295283

Mislevy, R. J. & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing* (Bericht Nr. RR88-48-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Bericht Nr. RR-96-30-0NR). Princeton, NJ: Educational Testing Service.

Mittelhaëuser, M.-A., Béguin, A. A. & Sijtsma, K. (2011). *Comparing the effectiveness of*

different linking designs: The internal anchor versus the external anchor and pre-test data (Measurement and Research Department Reports Nr. 2011-1). Arnhem, NL: Cito.

Mittelhaëuser, M.-A., Béguin, A. A. & Sijtsma, K. (2013). Modeling differences in test-taking motivation: Exploring the usefulness of the mixture Rasch model and person-fit statistics. In R. E. Millsap, L. A. van der Ark, D. M. Bolt & C. M. Woods (Hrsg.), *New developments in quantitative psychology* (S. 357-370). New York, NY: Springer. doi: 0.1007/978-1-4614-9348-8_ _23

Mittelhaëuser, M.-A., Béguin, A. A. & Sijtsma, K. (2015a). The effect of differential motivation on IRT linking. *Journal of Educational Measurement*, *52*, 339-358. doi: 10.1111/jedm.12080

Mittelhaëuser, M.-A., Béguin, A. A. & Sijtsma, K. (2015b). Selecting a data collection design for linking in educational measurement: Taking differential motivation into account. In R. E. Millsap, D. M. Bolt, L. A. van der Ark & W.-C. Wang (Hrsg.), *Quantitative psychology research: The 78th annual meeting of the psychometric society* (S. 181-193). New York, NY: Springer. doi: 10.1007/978-3-319-07503-7_ _11

Mroch, A. A. & Bolt, D. M. (2006). *An IRT-based response likelihood approach for addressing test speededness.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME). San Francisco, CA.

Müller, A. G. & Stanat, P. (2006). Schulischer Erfolg von Schülerinnen und Schülern mit Migrationshintergrund: Analysen zur Situation von Zuwanderern aus der ehemaligen Sowjetunion und aus der Türkei. In J. Baumert, P. Stanat & R. Watermann (Hrsg.), *Herkunftsbedingte Disparitäten im Bildungswesen: Vertiefende Analysen im Rahmen von PISA 2000* (S. 221-255). Wiesbaden: VS Verlag für Sozialwissenschaften.

Mullis, I. V., Martin, M. O., Foy, P. & Drucker, K. T. (2012). *PIRLS 2011 international results in reading.* Chestnut Hill, MA: International Association for the Evaluation

of Educational Achievement (IEA).

Mullis, I. V., Martin, M. O. & Loveless, T. (2016). *20 years of TIMSS: International trends in mathematics and science achievement, curriculum, and instruction*. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement (IEA).

Muthén, B., Asparouhov, T., Hunter, A. M. & Leuchter, A. F. (2011). Growth modeling with nonignorable dropout: Alternative analyses of the STAR∗D antidepressant trial. *Psychological Methods*, *16*, 17-33. doi: 10.1037/a0022634

Muthén, B. & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics*, *30*, 27-58. doi: 10.3102/10769986030001027

Muthén, L. K. & Muthén, B. O. (1998-2012). *Mplus user's guide* (7. Aufl.). Los Angeles, CA: Muthén & Muthén.

Nagy, G., Haag, N., Lüdtke, O. & Köller, O. (2017). Längsschnittskalierung der Tests zur überprüfung des Erreichens der Bildungsstandards der Sekundarstufe I im PISA-Längsschnitt 2012/2013. *Zeitschrift für Erziehungswissenschaft*, *20*, 259-286. doi: 10.1007/s11618-017-0755-1

Nagy, G., Lüdtke, O. & Köller, O. (2016). Modeling test context effects in longitudinal achievement data: Examining position effects in the longitudinal German PISA 2012 assessment. *Psychological Test and Assessment Modeling*, *58*, 641-670.

Nagy, G., Lüdtke, O., Köller, O. & Heine, J.-H. (2017). IRT-Skalierung der Tests im PISA-Längsschnitt 2012/2013: Auswirkungen von Testkontexteffekten auf die Zuwachsschätzung. *Zeitschrift für Erziehungswissenschaft*, *20*, 229-258. doi: 10.1007/s11618-017-0749-z

Nagy, G., Nagengast, B., Frey, A., Becker, M. & Rose, N. (2016). Itempositionseffekte in Large-Scale-Assessments. In Bundesministerium für Bildung und Forschung (Hrsg.), *Forschungsvorhaben in Ankopplung an Large-Scale-Assessments* (Bd. 44, S. 121-139). Berlin: BMBF.

Nagy, G., Retelsdorf, J., Goldhammer, F., Schiepe-Tiska, A. & Lüdtke, O. (2017). Veränderungen der Lesekompetenz von der 9. zur 10. Klasse: Differenzielle Entwicklungen in Abhängigkeit der Schulform, des Geschlechts und des soziodemografischen Hintergrunds? *Zeitschrift für Erziehungswissenschaft*, *20*, 177-203. doi: 10.1007/s11618-017-0747-1

Nagy, G. P. (1986). Validity of CTBS from an analysis of speededness and item response patterns. *Canadian Journal of Education*, *11*, 536-556.

NCES. (2016). *NAEP technical documentation.* National Center for Education Statistics. Zugriff auf `https://nces.ed.gov/nationsreportcard/tdw/` (Zugriff am 16.09.2017)

Nylund, K. L., Asparouhov, T. & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling*, *14*, 535-569. doi: 10.1080/10705510701575396

Nylund-Gibson, K. & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling*, *23*, 782-797. doi: 10.1080/10705511.2016.1221313

OECD. (2009). *PISA 2006 technical report.* Paris, F: OECD Publishing. (Organisation for Economic Co-operation and Development)

OECD. (2012). *PISA 2009 technical report.* Paris, F: OECD Publishing. (Organisation for Economic Co-operation and Development)

OECD. (2014). *PISA 2012 technical report.* Paris, F: OECD Publishing. (Organisation for Economic Co-operation and Development) doi: 10.1787/9789264167872-en

OECD. (2017). *PISA 2015 technical report.* Paris, F: OECD Publishing. (Organisation for Economic Co-operation and Development)

O'Muircheartaigh, C. & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical*

Society Series A (Statistics in Society), *162*, 177-194.

Oshima, T. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 200-219. doi: 10.1111/j.1745-3984.1994.tb00443.x

Penk, C., Pöhlmann, C. & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-scale Assessments in Education*, *2*, 1-17. doi: 10.1186/s40536-014-0005-4

Penk, C. & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, *29*, 55-79. doi: 10.1007/s11092-016-9248-7

Pietsch, M. (2011). Fehlende Daten bei Unterrichtsbeobachtungen: Eine Sensitivitätsanalyse anhand von Daten der Schulinspektion Hamburg. *Empirische Pädagogik*, *25*, 47-87.

Pietsch, M. & Stubbe, T. C. (2007). Inequality in the transition from primary to secondary school: school choices and educational disparities in Germany. *European Educational Research Journal*, *6*, 424-445. doi: 10.2304/eerj.2007.6.4.424

Pohl, S. & Carstensen, C. H. (2012). *NEPS technical report – Scaling the data of the competence tests* (NEPS Working Paper Nr. 14). Bamberg: National Educational Panel Study (NEPS).

Pohl, S. & Carstensen, C. H. (2013). Scaling of competence tests in the national educational panel study – Many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*, 189-216.

Pohl, S., Gräfe, L. & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74*, 423-452.

doi: 10.1177/0013164413504926

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (Hrsg.). (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland.* Münster: Waxmann.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Erweiterte Aufl.). Chicago, IL: The University of Chicago Press.

Retelsdorf, J., Becker, M., Köller, O. & Möller, J. (2012). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. *British Journal of Educational Psychology*, *82*, 647-671. doi: 10.1111/j.2044-8279 .2011.02051.x

Retelsdorf, J., Lindner, C., Nickolaus, R., Winther, E. & Köller, O. (2013). Forschungsdesiderate und Perspektiven – Ausblick auf ein Projekt zur Untersuchung mathematisch-naturwissenschaftlicher Kompetenzen in der beruflichen Erstausbildung (ManKobE). *Zeitschrift für Berufs- und Wirtschaftspädagogik, Beiheft 26, 227-234.*, *Beiheft 26*, 227-234.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, *16*, 261-270. doi: 10.1111/j.1745-3984.1979.tb00107.x

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards in Deutsch und Mathematik – Leistungsmessung in der Grundschule* (S. 42-106). Weinheim: Beltz.

Robitzsch, A. (2016). *Essays zu methodischen Herausforderungen im Large-Scale Assessment* (Unveröffentlichte Dissertation). Humboldt Universität zu Berlin.

Rohwer, G. (2013). *Making sense of missing answers in competence tests* (NEPS Working Paper Nr. 30). Bamberg: National Educational Panel Study (NEPS).

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* (Unveröffentlichte Dissertation). Friedrich-Schiller-Universität Jena.

Rose, N., von Davier, M. & Nagengast, B. (2015). Commonalities and differences in IRT-based methods for nonignorable item nonresponses. *Psychological Test and Assessment Modeling*, *57*, 472-498.

Rose, N., von Davier, M. & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, *82*, 795-819. doi: 10.1007/s11336-016-9544-7

Rose, N., von Davier, M. & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (Bericht Nr. RR-10-11). Princeton, NJ: Educational Testing Service.

Rost, J. (1988). *Quantitative und qualitative probabilistische Testtheorie.* Bern: Hans Huber.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271-282. doi: 10.1177/014662169001400305

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

Rost, J. & Langeheine, R. (1997). A guide through latent structure models for categorical data. In J. Rost & R. Langeheine (Hrsg.), *Applications of latent trait and latent class models in the social seiences* (S. 13-37). Münster: Waxmann.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592. doi: 10.1093/biomet/63.3.581

Salentin, K. (2014). Sampling the ethnic minority population in Germany. The background to "migration background". *methods, data, analyses*, *8*, 25-52. doi: 10.12758/mda.2014.002

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177. doi: 10.1037//1082-989X.7.2.147

Scherer, R., Greiff, S. & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*, 37-50. doi: 10.1016/

j.intell.2014.10.003

Schiepe-Tiska, A., Rönnebeck, S., Heitmann, P., Schöps, K., Prenzel, M. & Nagy, G. (2017). Die Veränderung der naturwissenschaftlichen Kompetenz von der 9. zur 10. Klasse bei PISA und den Bildungsstandards unter Berücksichtigung geschlechts- und schulart-spezifischer Unterschiede sowie der Zusammensetzung der Schülerschaft. *Zeitschrift für Erziehungswissenschaft*, *20*, 151-176. doi: 10.1007/s11618-017-0754-2

Schmitt, A. P. & Bleistein, C. A. (1987). *Factors affecting differential item functioning for Black examinees on Scholastic Aptitude Test analogy items* (Bericht Nr. 87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A. P. & Crone, C. R. (1991). *Alternative mathematical aptitude item types: DIF issues* (Bericht Nr. RR-91-42). Princeton, NJ: Educational Testing Service.

Schmitt, A. P. & Dorans, N. J. (1988). *Differetial item functioning for minority examinees on the SAT* (Bericht Nr. RR-88-32). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., Dorans, N. J., Crone, C. R. & Maneckshana, B. T. (1991). *Differential speededness and item omit patterns on the SAT* (Bericht Nr. RR-91-50). Princeton, NJ: Educational Testing Service.

Schneider, B., Martinez, S. & Owens, A. (2006). Barriers to educational opportunities for Hispanics in the United States. In M. Tienda & F. Mitchell (Hrsg.), *Hispanics and the future of America* (S. 179-227). Washington DC: National Academics Press.

Schnipke, D. L. & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213-232. doi: 10.1111/j.1745-3984.1997.tb00516.x

Schnipke, D. L. & Scrams, D. J. (2002). Exploring the issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. Fremer & W. C. Ward (Hrsg.), *Computer-bases testing: Building the foundation of future assessments* (S. 237-266). Mahwah, NJ: Lawrence Erlbaum.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464.

Sessoms, J. & Finney, S. J. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing, 15: 356-388, 2015*, *15*, 356-388. doi: 10.1080/15305058.2015.1034866

Shealy, R. T. & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Hrsg.), *Differential item functioning* (S. 197-239). Mahwah, NJ: Lawrence Erlbaum.

Singer, J. D. & Willett, J. B. (1993). It's about time: using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, *18*, 155-195. doi: 10.2307/1165085

Sinharay, S. & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, *33*, 23-35. doi: 10.1111/emip.12024

Sireci, S. G., Han, K. T. & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment*, *13*, 108-131. doi: 10.1080/10627190802394255

Solga, H. (2005). *Ohne Abschluss in die Bildungsgesellschaft: Die Erwerbschancen gering qualifizierter Personen aus soziologischer und ökonomischer Perspektive.* Opladen: Barbara Budrich. doi: 10.3224/93809407

Stanat, P. & Artelt, C. (2009). Schulleistungsuntersuchungen. In S. Blömeke, T. Bohl, L. Haag, G. Lang-Wojtasik & W. Sacher (Hrsg.), *Handbuch Schule: Theorie – Organisation – Entwicklung* (S. 119-125). Bad Heilbrunn: Julius Klinkhardt.

Stanat, P., Rauch, D. & Segeritz, M. (2010). Schülerinnen und Schüler mit Migrationshintergrund. In E. Klieme et al. (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt* (S. 200-230). Münster: Waxmann.

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S. & Haag, N. (Hrsg.). (2017). *IQB-Bildungstrend 2016.* Münster: Waxmann.

Stenlund, T., Eklöf, H. & Lyrén, P.-E. (2016). Group differences in test-taking behaviour: An example from a high-stakes testing program. *Assessment in Education: Principles, Policy & Practice*. doi: 10.1080/0969594X.2016.1142935

Suh, Y., Cho, S.-J. & Wollack, J. A. (2012). A comparison of item calibration procedure in the presence of test speededness. *Journal of Educational Measurement*, *49*, 285-311. doi: 10.1111/j.1745-3984.2012.00176.x

Suh, Y., Kang, T., Wollack, J. A. & Kim, S.-Y. (2006). *A comparison of test scoring methods in the presence of test speededness.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME). San Francisco, CA.

Sundre, D. L. (1999). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the Annual Meeting of the American Educational Research Association (AERA). Montreal, Canada.

Tate, R. (2002). Test dimensionality. In G. Tindal & T. M. Haladyna (Hrsg.), *Large-scale assessment programs for all students* (S. 181-211). Mahwah, NJ: Lawrence Erlbaum.

The College Board. (o. J.). *SAT suite of assessments.* New York, NY. Zugriff auf `https://collegereadiness.collegeboard.org/sat` (Zugriff am 16.09.2017)

Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich & P. W. Holland (Hrsg.), *Linking and aligning scores and scales* (S. 287-312). New York, NY: Springer.

van Barneveld, C. (2003). The effects of examinee motivation on multiple-choice item parameter estimates. *The Alberta Journal of Educational Research*, *XLIX*, 277-289.

van Barneveld, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement*, *31*, 31-46. doi: 10.1177/0146621606286206

van der Linden, W. J. (2007a). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287-308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2007b). *Test design and speededness* (Bericht Nr. 07-01). Newtown, PA: Law School Admission Council.

van der Linden, W. J. (2011). Setting time limits on tests. *Applied Psychological Measurement*, *35*, 183-199. doi: 10.1177/0146621610391648

van der Linden, W. J., Scrams, D. J. & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195-210. doi: 10.1177/01466219922031329

Verhelst, N. D., Glas, C. A. W. & de Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 123-138). New York, NY: Springer.

Vermunt, J. K. & Magidson, J. (2005). *Latent GOLD 4.0 user's guide.* Belmont, MA: Statistical Innovations Inc.

Vidotto, D. (2018). *Bayesian latent class models for the multiple imputation of cross-sectional, multilevel and longitudinal categorical data* (Unveröffentlichte Dissertation). Tilburg University, NL.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287-307. doi: 10 .1348/000711007X193957

von Davier, M. & Yamamoto, K. (2007). Mixture-distribution and hybrid Rasch models. In M. von Davier & C. H. Carstensen (Hrsg.), *Multivariate and mixture distribution rasch models: Extensions and applications* (S. 99-115). New York, NY: Springer.

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications.* New York, NY: Cambridge University Press.

Walter, O. & Rost, J. (2011). Psychometrische Grundlagen von Large Scale Assessments.

In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Methoden der psychologischen Diagnostik* (S. 88-150). Göttingen: Hogrefe.

Wang, A. (2011). *A mixture cross-classification IRT model for test speededness* (Unveröffentlichte Dissertation). University of Georgia, Athens, GA.

Weirich, S. (2015). *Kontexteffekte in Large-scale Assessments (Context effects in large-scale assessments)* (Unveröffentlichte Dissertation). Humboldt Universität zu Berlin.

Wild, C. L., Durso, R. & Rubin, D. B. (1982). Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement*, *19*, 19-28. doi: 10.1111/j.1745-3984.1982.tb00111.x

Wise, L. L. (1996). *A persistence model of motivation and test performance.* Paper presented at the annual convention of the American Educational Research Association (AERA). New York, NY.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*, 95-11. doi: 10.1207/s15324818ame1902_2

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

Wise, S. L. & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*, 27-41. doi: 10.1080/10627191003673216

Wise, S. L., Kingsbury, G. G., Thomason, J. & Kong, X. (2004). *An investigation of motivation filtering in a statewide achievement testing program.* Paper Presented at the Annual meeting of the National Council of Measurement in Education (NCME). San Diego, CA.

Wise, S. L. & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163-183. doi: 10

.1207/s15324818ame1802_2

Wise, S. L., Pastor, D. A. & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*, 185-205. doi: 10.1080/08957340902754650

Wolf, L. F. & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*, 227-242. doi: 10.1207/s15324818ame0803_3

Wolf, L. F., Smith, J. K. & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education*, *8*, 341-351. doi: 10.1207/s15324818ame0804_4

Wollack, J. A., Cohen, A. S. & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement*, *40*, 307-330. doi: 10.1111/j.1745-3984.2003.tb01149.x

Wollack, J. A., Suh, Y. & Bolt, D. M. (2007). *Using the testlet model to mitigate test speededness effects.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME). Chicago, IL.

Wu, M. C. (2010). Measurement, sampling, and equating error in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*, 15-27. doi: 10.1111/j.1745-3992.2010.00190.x

Wu, M. C. & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44*, 175-188. doi: 10.2307/2531905

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the hybrid model* (Bericht Nr. TR-95-2). Princeton, NJ: Educational Testing Service.

Yamamoto, K. & Everson, H. (1995). *Modeling the mixture of IRT and pattern responses by*

a modified hybrid model (Bericht Nr. RR-95-16). Princeton, NJ: Educational Testing Service.

Yamamoto, K. & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost & R. Langeheine (Hrsg.), *Applications of latent trait and latent class models in the social sciences* (S. 89-98). New York, NY: Waxmann.

Yang, X. (2007). Methods of identifying individual guessers from item response data. *Educational and Psychological Measurement*, *67*, 745-764. doi: 10.1177/0013164406296978

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameter for two latent trait models. *Journal of Educational Measurement*, *17*, 297-311. doi: 10.1111/j.1745-3984.1980.tb00833.x

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213. doi: 10.1111/j.1745-3984.1993.tb00423.x

Yen, W. M. & Fitzpatrick, A. R. (2006). Item response theory. In R. Brennan (Hrsg.), *Educational measurement* (4. Aufl., S. 111-153). Westport, CT: Praeger.

Lebenslauf

Marit Kristine List, Dipl. Psych., B. Sc. (Kognitionswissenschaft)

Akademischer Werdegang

2013–2018	Promotionsstudium (Psychologie), Christian-Albrechts-Universität zu Kiel
2006–2013	Bachelorstudium (Kognitionswissenschaft), Universität Osnabrück
2006–2011	Diplomstudium (Psychologie), Universität Osnabrück
2005–2006	Gaststudium, Lawrence University, Appleton, WI, USA
2004–2005	Diplomstudium (Psychologie), Universität Hamburg
2003	Abitur, Gymnasium Carolinum, Osnabrück

Berufliche Tätigkeiten

seit 2018	Wissenschaftliche Mitarbeiterin, CESR – Center for Environmental Systems Research an der Universität Kassel
2013–2018	Wissenschaftliche Mitarbeiterin und Doktorandin, IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Kiel
2011–2012	Wissenschaftliche Mitarbeiterin, nifbe e.V. – Niedersächsisches Institut für frühkindliche Bildung und Entwicklung an der Universität Osnabrück