Nagy, Gabriel; Frey, Andreas; Nagengast, Benjamin; Becker, Michael; Rose, Norman

Itempositionseffekte in Large-Scale-Assessments

Bundesministerium für Bildung und Forschung [Hrsg.]: Forschungsvorhaben in Ankopplung an Large-Scale-Assessments. Berlin : Bundesministerium für Bildung und Forschung 2016, S. 121-139. - (Bildungsforschung; 44)

Gabriel Nagy, Benjamin Nagengast, Andreas Frey, Michael Becker, Norman Rose

Itempositionseffekte in Large-Scale-Assessments

Das vorliegende Projekt widmet sich einem spezifischen Testkontexteffekt, der als (Item-)Positionseffekt bekannt ist und die systematische Variation von statistischen Eigenschaften von Einzelitems in Abhängigkeit ihrer Darbietungsposition in einem Leistungstest bezeichnet. Im Fokus der Betrachtung stehen individuelle Unterschiede in der Höhe von Positionseffekten, deren Korrelate und die Auswirkungen der Vernachlässigung von Positionseffekten insbesondere in Large-Scale-Assessments. Die auf Grundlage exitierender Datensätze von Large-Scale-Assessments durchgeführten Analysen dokumentieren systematische Zusammenhänge von Positionseffekten mit Merkmalen der Schülerinnen und Schüler, der Einzelschulen und der Erhebungszeitpunkte. Die Ergebnisse zeigen ferner, dass Positionseffekte zu ignorieren zu verzerrten Ergebnissen hinsichtlich Leistungsunterschieden, Korrelaten von Testleistungen und Leistungszuwächsen führen kann.

1 Hintergrund

Die Erfassung schülerseitiger Kompetenzen ist das Hauptanliegen nahezu aller bildungswissenschaftlicher Large-Scale-Assessments. Diese im Deutschen auch als Schulleistungsstudien bezeichneteten groß angelegten empirischen Untersuchungen sollen einen Einblick in die Verteilung der Kompetenzausprägungen und deren Entwicklung in einer oder mehrere Populationen von Schülerinnen und Schülern geben. Um dieses Ziel zu erreichen, wurden sophistizierte Messmodelle und Prozeduren entwickelt, die eine unverzerrte und gleichermaßen reliable Schätzung der Populationsverteilung der Kompetenzen ermöglichen sollen (z. B. Wu, 2005). Typische Ansätze der Kompetenzmessung sind in der Tradition der Item-Response-Theorie (IRT) verwurzelt. Ziel dieser Verfahren ist, die den beobachteten Itemantworten zugrunde liegenden Kompetenzausprägungen abzuschätzen.

Eine zentrale Idee des IRT-Ansatzes ist, dass die beobachteten Itemantworten eine Funktion feststehender Itemeigenschaften (z. B. Schwierigkeiten) und Personenmerkmale (d. h. Kompetenzausprägungen) sind. Bisherige Untersuchungen liefern aber eine Fülle von Belegen dafür, dass diese Annahme in vielen Schulleistungsstudien nicht erfüllt ist. Ein robuster Befund ist, dass die Wahrscheinlichkeit einer korrekten Itemantwort von der Position des Items in der dargebotenen Sequenz von Testitems abhängt. Je näher ein Item zum Testende hin positioniert wird, desto geringer fällt die Lösungswahrscheinlichkeit des Items typischerweise aus (Leary & Dorans, 1985). Derartige Itempositionseffekte können als eine Stör-

quelle verstanden werden, die sich auf die Ergebnisse der Leistungsmessung auswirkt.

Die Implikationen von Itempositionseffekten für die Validität der Rückschlüsse in Large-Scale-Assessments hängen maßgeblich von den Eigenschaften der Itempositionseffekte ab. Itempositionseffekte wurden lange Zeit als feststehende Effekte betrachtet, deren Ausprägung nicht zwischen Individuen, Schulen und Messzeitpunkten variiert (z. B. Meyers, Miller & Way, 2009). Insofern diese Sichtweise zutrifft, wirken sich Itempositionseffekte nicht zwingend auf Leistungsvergleiche zwischen Schülerinnen und Schülern, Gruppen von Schülerinnen und Schülern und Messzeitpunkten aus, da diese Effekte mittels eines geeigneten Testheftdesigns (Frey, Hartig & Rupp, 2009) ausbalanciert bzw. konstant gehalten werden können (Johnson, 1990). Neuere Arbeiten liefern jedoch Belege dafür, dass die Höhe von Itempositionseffekten zwischen Personen (Debeer & Janssen, 2013; Hecht, Weirich, Siegle & Frey, 2015), Schulen (Debeer, Buchholz, Hartig & Janssen, 2014) und Messzeitpunkten (Nagy, Lüdtke, Köller, Heine & Mang, 2015) variieren kann. Derartige Störeinflüsse können nicht durch das Testheftdesign kontrolliert werden, da die Höhe des Itempositionseffekts keine alleinige Funktion der verwendeten Testform ist.

Ausgehend von der Feststellung, dass die Ausprägung von Itempositionseffekten zwischen Personen variiert, wurden diese als Indikatoren der individuellen Bearbeitungspersistenz in einem Leistungstest vorgeschlagen (Debeer et al., 2014). Diese Sichtweise impliziert, dass die in Large-Scale-Assessments verwendeten Testwerte neben der zu erfassenden Kompetenzausprägung auch von der individuellen Bearbeitungspersistenz abhängen. Damit ergeben sich direkte Konsequenzen für die Validität von Rückschlüssen, die auf Grundlage von Testwerten getroffen werden. Diese betreffen unter anderem die Abschätzung (1) der Leistungsheterogenität in einer (Teil-)Population, (2) der Zusammenhänge individueller Kompetenzausprägungen mit anderen Merkmalen und (3) von Kompetenzzuwächsen über einen vorgegebenen Beschulungszeitraum.

Interindividuell variierende Itempositionseffekte führen zu einer Überschätzung der Leistungsheterogenität, da sie die Variabilität von Testwerten künstlich erhöhen. Ebenso gilt, dass Korrelationen zwischen Testwerten und Außenkriterien ein verzerrtes Abbild der entsprechenden Kriteriumskorrelationen der Kompetenzen darstellen, da die berechneten Zusammenhänge auch von der Korrelation der Bearbeitungspersistenz mit den untersuchten Kovariaten abhängen. Schließlich können sich Itempositionseffekte auf die Abschätzung von Leistungszuwächsen auswirken, da sich die Bearbeitungspersistenz der getesteten Schülerinnen und Schüler zwischen den Messzeitpunkten verändern kann.

Die aufgezählten Implikationen von Itempositionseffekten ergeben sich direkt aus deren Konzeption als variable Störgrößen. Bis dato finden sich aber kaum Arbeiten aus dem Schulleistungsbereich, die sich der Frage der Korrelate und der zeitlichen Veränderung von Itempositionseffekten gewidmet haben. Im vorliegenden Projekt werden diese Forschungslücken aufgegriffen und exemplarisch anhand existierender Datensätze untersucht.

2 Zielstellungen des Projekts

Das vorliegende Projekt gliedert sich in vier Teilstudien auf, die sich der Untersuchung unterschiedlicher Aspekte von Itempositionseffekten widmen. Studie 1 fokussiert auf die Erfassung individueller Unterschiede in Itempositionseffekten und deren Korrelate mittels eines mehrdimensionalen IRT-Modells. In Studie 2 werden Positionseffekte in der nationalen Erweiterung der PISA-2006-Studie untersucht. Konkret werden Prädiktoren von Positionseffekten auf der Schüler- und Schulebene betrachtet. In der dritten Studie werden Itempositionseffekte im Längsschnitt untersucht. Im Fokus steht die mittlere Veränderung sowie die Rangstabilität von Itempositionseffekten. Die vierte Studie widmet sich schließlich der Frage, inwieweit Itempositionseffekte mit anderen Testkontexteffekten (Brennan, 1992) interagieren.

Die Erfassung individueller Unterschiede in Itempositionseffekten setzt erstens die Verfügbarkeit von Testheftdesigns voraus, in denen die gleichen Items unterschiedlichen Schülerinnen und Schülern an unterschiedlichen Positionen eines Tests dargeboten werden. Zweitens setzen die verwendeten Verfahren eine ausreichend große Stichprobe voraus, da sie sich in der Regel mehrdimensionaler IRT-Verfahren bedienen (Debeer & Janssen, 2013). Diese Vorgaben implizieren, dass viele derzeit frei nutzbare Large-Scale-Datensätze nicht für die hier anvisierten Arbeiten in Betracht kamen, da sie keine ausreichende Variation der Positionen einzelner Items bieten. Die Durchführung einer eigens geplanten Primärdatenerhebung hat zwar einige Vorteile, wurde aber vor dem Hintergrund des zum Zeitpunkt der Projektplanung herrschenden Wissensstands als zu kostenintensiv eingeschätzt.

Die ausgewählten Datensätze ermöglichen die Modellierung individueller Unterschiede in Itempositionseffekten und erlauben die Behandlung unterschiedlicher Aspekte des Gegenstandbereichs, die sowohl grundlagenwissenschaftlich als auch anwendungsorientiert ausgerichtet sind. So liefert der Datensatz der TRAIN-Studie individuelle Hintergrundvariablen, die aus einer theoretischen Perspektive als zentrale Determinaten der Bearbeitungspersistenz betrachtet werden können. Dieser Datensatz ermöglicht zudem die längsschnittliche Untersuchung von Itempositionseffekten. Der PISA-2006-Datensatz ist von seiner Struktur her prototypisch für viele aktuelle Large-Scale-Assessments und ermöglicht es abzuschätzen, inwieweit Positionseffekte mit Hintergrundvariablen kovariieren, die im Fokus traditioneller Schulleistungsstudien stehen (Schulform, Geschlecht und Merkmale des familiären Hintergrunds). Ein weiterer Datensatz aus dem Projekt MaK-adapt (Messung allgemeiner Kompetenzenadaptiv), mit Daten von drei computerbasierten Tests, bietet schließlich die seltene Möglichkeit, die Interaktion von Itempositionseffekten mit anderen Testkontexteffekten (Effekte der Domänenabfolge) zu untersuchen.

Im vorliegenden Beitrag werden die zentralen Befunde der ersten drei Studien zusammengefasst. Die vorgenommene Fokussierung ist insofern konsistent, da sich die ausgewählten Studien eng an der Thematik individueller Unterschiede in Itempositionseffekten ausrichten, während diese Perspektive in Studie 4 eine untergeordnete Rolle einnimmt. In diesem Überblicksbeitrag beschränken wir uns auf die Darstellung der inhaltlich-substanzwissenschaftlichen Ergebnisse. Verzichtet wird auf die Darstellung der verwendeten mathematischen Modelle. Um die Nachvoll-

ziehbarkeit der Ergebnisse zu gewährleisten, setzen wir, wann immer möglich, grafische Hilfsmittel ein.

3 Studie 1: Itempositionseffekte in einem Leseverständnistest

Die erste Studie (Nagy, Rose & Trautwein, 2013) verfolgte zwei Ziele. Aus methodischer Perspektive wurde ein IRT-Modell vorgestellt, das zwei Arten von Positionseffekten umfasst: zum einen Positionseffekte, die auf die Itemschwierigkeiten einwirken und zwischen Personen variieren können, und zum anderen Itempositionseffekte, die auf die Diskriminationsleistung der Items einwirken und zu einer Reduktion der Sensitivität der Testitems für die zur Lösung benötigte Kompetenz führen. Das verwendete Modell ist eine Weiterentwicklung des Verfahrens von Debeer und Janssen (2013).

Das zweite Ziel der Studie war es, Zusammenhänge zwischen Positionseffekten und schülerseitigen motivationalen und kognitiven Merkmalen zu untersuchen, die aus einer theoretischen Perspektive als Determinaten der Testbearbeitungspersistenz in Betracht kommen. Die Auswertungen liefern wichtige Hinweise für die inhaltliche Bedeutung der untersuchten Itempositionseffekte und zeigen zudem, inwieweit sich Itempositionseffekte verzerrend auf die ermittelten Konstruktzusammenhänge auswirken können. Die in dieser Studie anvisierten Forschungsfragen wurden anhand eines Leseverständnistests, der in der fünften Jahrgangsstufe in der TRAIN-Studie (Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen) eingesetzt wurde, untersucht (N = 2.830 Schülerinnen und Schüler an Haupt-, Real- und Mittelschulen, 53,6 Prozent männlich, die jeweils 30 bis 32 Items bearbeitet haben). Zwei Kovariaten wurden zur Prädiktion der Testleistung und der Bearbeitungspersistenz herangezogen, nämlich eine Selbstberichtsskala zur Erfassung der Freude am Lesen und ein Instrument zur Messung der Decodiergeschwindigkeit. Als Hypothesen wurden angenommen, dass Schülerinnen und Schüler, die für den Bereich Lesen ein höheres Interesse berichten, eine höhere Persistenz bei der Bearbeitung von Textaufgaben aufweisen sowie hinsichtlich der Decodiergeschwindigkeit, dass sich diese förderlich auf die Bearbeitungspersistenz auswirkt, da der Prozess des Lesens für schnelle Decodierer weniger ermüdend ist.

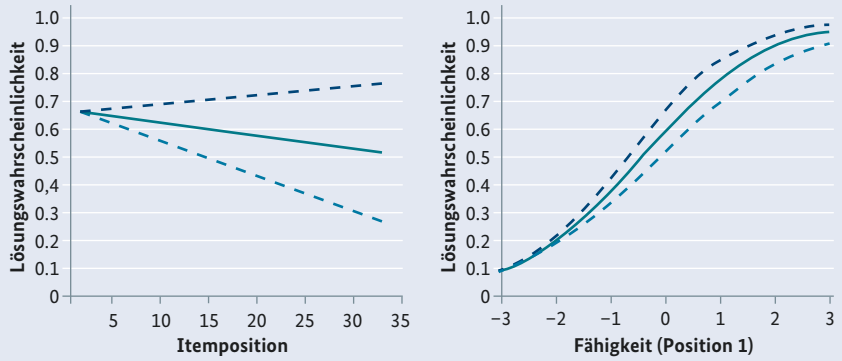
3.1 Zentrale Ergebnisse

Hinsichtlich der Positionseffekte wurde festgestellt, dass die Wahrscheinlichkeit korrekter Antworten mit der Itemposition assoziiert ist, wobei die Höhe der Abnahme zwischen Personen variiert. Ebenso wurde ein Rückgang der Sensitivität der Testitems zur Erfassung der im Fokus stehenden fokalen Kompetenz (d. h. des Leseverständnisses) ermittelt. Die entsprechenden Ergebnisse sind in Abbildung 1 dargestellt. Die linke Teilabbildung stellt den Rückgang der Lösungswahrscheinlichkeit eines prototypischen Items (d. h. mittlere Itemschwierigkeit und Itemdiskrimination) dar. Die Abbildung dokumentiert eine kontinuierliche Abnahme der Lösungswahrscheinlichkeit in Abhängigkeit der Itemposition. Individuelle Unterschiede in

der Testbearbeitungspersistenz manifestieren sich in der Variation der Leistungsabnahme zwischen Personen. Aus der Abbildung geht hervor, dass ein Teil der Schülerinnen und Schüler keinen Leistungsrückgang während der Testbearbeitung aufweisen, während für andere der Rückgang besonders stark akzentuiert ist.

In der rechten Teilabbildung 1 ist die Veränderung der Itemcharakteristikkurve eines prototypischen Items dargestellt. Die Abbildung verdeutlicht, dass der ogivenförmige Zusammenhang zwischen der Lösungswahrscheinlichkeit eines Items (y-Achse) und der individuellen Kompetenzausprägung (x-Achse) über die Itempositionen hinweg flacher wird. Dieser Befund indiziert, dass die zum Testende hin eingesetzten Items weniger sensitiv für individuelle Unterschiede in der zugrunde liegenden Kompetenz sind, wobei die Sensitivität für Unterschiede in der Testbearbeitungspersistenz zunimmt (ohne Abbildung).

Abb. 1: Linke Teilabbildung: Lösungswahrscheinlichkeit eines prototypischen Items in Abhängigkeit der Itemposition. Vorhersage für die mittlere Ausprägung des Positionseffekts und des Wertebereichs zwischen dem 10. und 90. Percentil der Verteilung des Positionseffekts. Rechte Teilabbildung: Itemchararakteristikkurven für ein prototyisches Item zu Beginn (obere gebrochene Linie), in der Mitte (durchgezogene Linie) und am Ende des Tests (untere gebrochene Linie)

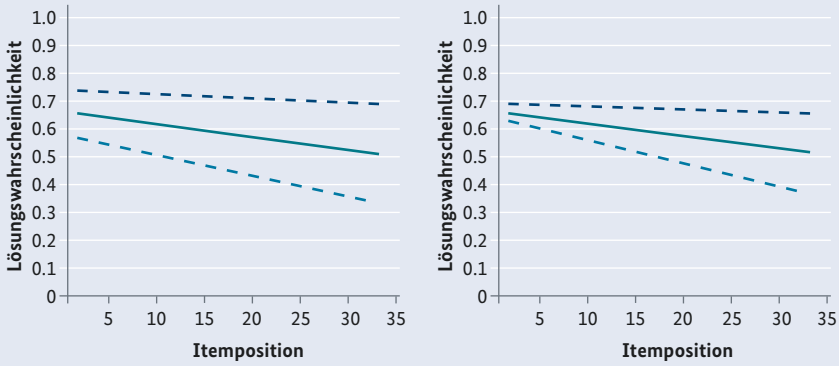


Das IRT-Modell wurde in einem zweiten Schritt um die Kovariaten Lesefreude und Decodiergeschwindigkeit erweitert. Dieses Modell ermöglicht es, die Zusammenhänge zwischen den Kovariaten und den Positionseffekten darzustellen. Es beantwortet somit die Frage, ob individuelle Unterschiede in der Testbearbeitungspersistenz mit den betrachteten Kovariaten in erwarteter Weise zusammenhängen. Auf Grundlage der Modellparameter können zudem Zusammenhänge zwischen der fokalen Testleistung und den Kovariaten getrennt für jede mögliche Itemposition berechnet werden. Damit ist es möglich, die Außenkorrelationen von Testwerten in Abhängigkeit unterschiedlicher Referenzpositionen darzustellen.

In dieser Anwendung zeigte sich, dass beide Kovariaten signifikant positiv mit dem Itempositionseffekt korreliert waren ($r = .21$; $SE = 0.04$; $p < .01$ und $r = .23$; $SE = 0.05$; $p < .01$ für Freude am Lesen und Dekodiergeschwindigkeit). Die Ergebnisse indizieren, dass Schülerinnen und Schüler mit einer höheren Lesefreude und einer

günstigeren Decodiergeschwindigkeit eine höhere Bearbeitungspersistenz aufweisen. Eine Folge dieses Zusammenhangs ist, dass die Leistungsunterschiede zwischen Schülerinnen und Schülern mit hohen und geringen Werten auf den Kovariaten in Abhängigkeit der Position der Leseverständnisitems zunehmen. Dieser Effekt ist in Abbildung 2 veranschaulicht. Eine Konsequenz dieses Ergebnismusters ist, dass die Zusammenhänge zwischen den Kovariaten und den Testergebnissen in Abhängigkeit der Position ansteigen. Die jeweiligen Kriteriumskorrelationen decken je nach Itemposition einen Wertebereich von $r = .10$ bis $.24$ (Lesefreude) bzw. $r = .27$ bis $.35$ (Decodiergeschwindigkeit) ab.

Abb. 2: Vorhergesagte Lösungswahrscheinlichkeit eines prototypischen Items in Abhängigkeit der Kovariaten Decodiergeschwindigkeit (linke Teilabbildung) und Freude am Lesen (rechte Teilabbildung). Durchgezogene Linie: Stichprobenmittelwert; gebrochene Linien: Vorhersagen für das 10. und 90. Percentil der Verteilung der Kovariaten



Ein weiterer wichtiger Befund, der sich in dieser Anwendung zeigte, ist, dass die aufgrund eines herkömmlichen IRT-Modells (ohne Berücksichtigung der Itempositionseffekte) ermittelten Konstruktzusammenhänge den aufgrund der im erweiterten Modell hinsichtlich der mittleren Itemposition erwarteten Zusammenhängen entsprachen. Dieser Befund dokumentiert, dass die in den meisten Studien ermittelten Zusammenhänge als Komposita der Kriteriumszusammenhänge mit dem intendierten Konstrukt (hier Leseverständnis) und mit der Testbearbeitungspersistenz zu verstehen sind.

3.2 Zusammenfassung

Studie 1 liefert wichtige methodische, grundlagenwissenschaftliche und anwendungsrelevante Ergebnisse. Der methodische Beitrag besteht in der Entwicklung eines flexiblen mehrdimensionalen IRT-Modells, das Itempositionseffekte erfasst, die auf Itemschwierigkeiten und die Diskriminationsleistung von Items einwirken. Das Modell stellt somit eine sinnvolle Erweiterung rezenter IRT-Ansätze zur Erfassung von Itempositionseffekten dar (z. B. Debeer & Janssen, 2013).

Aus einer inhaltlichen Perspektive tragen die Befunde zur Klärung des Konstruktstatus von Itempositionseffekten bei. Deren Interpretation als Indikatoren der Testbearbeitungspersistenz setzt voraus, dass diese mit motivationalen und kognitiven Ressourcen assoziiert sind. Die Ergebnisse der Studie 1 untermauern die Sichtweise, dass Itempositionseffekte einen Aspekt der Testbearbeitungspersistenz darstellen.

Studie 1 dokumentiert zudem die Bedeutung von Itempositionseffekten für die Abschätzung von Konstruktzusammenhängen. Kriteriumskorrelationen von (IRT-skalierten) Testwerten werden typischerweise als Zusammenhänge zwischen individuellen Kompetenzen und anderen Variablen interpretiert. Die Abhängigkeit der ermittelten Zusammenhänge von der Itemposition dokumentiert jedoch die Gefahr, die aus einer automatischen Gleichsetzung von Kompetenzen und Testwerten resultiert. Insgesamt liefern unsere Befunde Hinweise für die Annahme, dass die in Large-Scale-Assessments typischerweise verwendeten Testwerte als Indikatoren der Kompetenz mit der Testbearbeitungspersistenz konfundiert sind und dies auch für die ermittelten Kriteriumskorrelationen gilt.

4 Studie 2: Positionseffekte in der PISA-2006-Studie

Gegenstand von Studie 2 (Nagengast, Nagy, Rose & Becker, 2015) war die Untersuchung von Positionseffekten in einem für Large-Scale-Assessments prototypischen Datensatz. Zu diesem Zweck wurde die nationale Ergänzung der PISA-2006-Studie gewählt. Ein zentrales Anliegen von Large-Scale-Assessments ist die Beschreibung der Kompetenzverteilung in unterschiedlichen Teilpopulationen von Schülerinnen und Schülern. Typische Beschreibungen richten sich auf Populationen, die anhand der Schulform, des Geschlechts, des Migrationshintergrunds oder des sozioökonomischen Hintergrunds definiert sind. In Studie 2 wurde untersucht, inwieweit Positionseffekte zu verzerrten Rückschlüssen über Leistungsunterschiede zwischen Teilpopulationen führen können.

Eine zweite Fragestellung, die in dieser Studie angegangen wurde, richtet sich auf die Variabilität von Positionseffekten auf der Ebene von Schulen. Bis heute liegt lediglich eine Untersuchung vor, die für den Bereich Leseverständnis zwischenschulische Unterschiede in der Ausprägung von Positionseffekten dokumentiert (Debeer et al., 2014). In Studie 2 untersuchten wir deshalb diese Fragstellung für alle in PISA getesteten Kompetenzbereiche (Naturwissenschaften, Mathematik und Lesen). Obwohl die entsprechenden Auswertungen deutliche Hinweise für zwischenschulische Unterschiede in Positionseffekten erbracht haben, haben wir uns aus Platzgründen dafür entschieden, diese im vorliegenden Überblicksbeitrag nicht zu berichten.[1]

1 Das Befundmuster lässt sich grob wie folgt zusammenfassen: Positionseffekte in allen untersuchten Tests wiesen eine statistisch signifikante Varianz auf Schulebene auf. Ein Teil dieser Variabilität konnte auf Kompositionseffekte der Schülerschaft zurückgeführt werden (Geschlecht, Migrationshintergrund, sozioökonomischer Hintergrund und selbst berichtete Testanstrengung). Darüber hinaus waren die auf der Schulebene vorliegenden Positionseffekte mit der Schulform (alle Leistungsbereiche), dem Anteil der Schülerschaft mit Migrationshintergrund

Als individuelle Prädiktoren der Testleistung und des Positionseffektes wurden in dieser Untersuchung folgende Variablen berücksichtigt: Geschlecht, Migrationshintergrund (mindestens ein Elternteil im Ausland geboren), sozioökonomischer familiärer Hintergrund und die zum Ende des Tests selbst berichtete Testanstrengung (dichotomisiert). In der nachfolgenden Darstellung beschränken wir uns aus Platzgründen auf die Effekte des Geschlechts, des sozioökonomischen Hintergrunds und der besuchten Schulform. Wir haben uns für diese Variablen entschieden, da diese im Fokus der Berichtserstattung aller Large-Scale-Assessments stehen.

4.1 Zentrale Ergebnisse

Die Auswertung der Daten geschah anhand eines Verfahrens, das mehrere Analyseschritte umfasste. Wir haben uns gegen eine direkte Modellierung von Itempositionseffekten entschieden, da deren Bestimmung anhand eines nicht linearen mehrebenenanalytischen IRT-Modells (Schülerinnen und Schüler geschachtelt in Schulen) sich als rechnerisch nicht handhabbar erwies. Stattdessen haben wir die Positionseffekte auf Ebene der Positionen der in PISA eingesetzten Itemcluster erfasst (d. h. Clusterpositionseffekte). In PISA werden die Testitems in Itemcluster zusammengefasst, deren Bearbeitungszeit jeweils ca. 30 Minuten beträgt. Jeweils vier Itemcluster werden anschließend zu einem Testheft zusammengefasst. Die Verteilung der Itemcluster geschieht dabei auf Basis eines balancierten unvollständigen Block-Designs, das unter anderem gewährleistet, dass jedes Itemcluster jeweils genau einmal an jeder der vier möglichen Clusterpositionen eines Testhefts dargeboten wird (Frey, Carstensen, Walter, Rönnebeck & Gomolka, 2008). Da die durch das Testheftdesign spezifizierten Testhefte randomisiert den Schülerinnen und Schülern zugeteilt wurden, können (Cluster-)Positionseffekte anhand der zwischen den Clusterpositionen vorliegenden Leistungsunterschiede ermittelt werden. Dieses Verfahren setzt voraus, dass für jede Kombination, die sich aus der Kreuzung der Itemcluster und der Clusterposition ergibt, Testwerte vorliegen. In Studie 2 wurden die Testwerte mittels getrennter Rasch-Skalierungen auf Grundlage der Plausible-Value-Technik (PV; Mislevy, Beaton, Kaplan & Sheehan, 1992) ermittelt. Das in PISA 2006 verwendete Testheftdesign ist in Tabelle 1 wiedergegeben.

Die für jede Domäne ermittelten PVs wurden anschließend mittels eines neu entwickelten multivariaten Mehrebenenmodells ausgewertet. In diesem Modell werden die verwendeten PVs in zwei Komponenten zerlegt, nämlich die mittlere Ausprägung der PVs über alle Itemcluster und alle Clusterpositionen sowie die positions- und clusterpezifischen Abweichungen von diesem Mittelwert. In seiner einfachsten Form erlaubt dieses Modell die Zerlegung der Variabilität der Testwerte in Leistungsmittelwerte und Positionseffekte getrennt für die Schüler- und Schulebene. Das Modell

(Mathematik) und dem Anteil der Schülerschaft mit geringer Testanstrengung (Naturwissenschaften und Lesen) assoziiert. Mit Ausnahme der Schulformeffekte lassen sich diese Effekte im Sinne von Kontexteffekten interpretieren, die darauf hinweisen, dass die Testbearbeitungspersistenz in Abhängigkeit der Zusammensetzung der Schülerschaft (nach Kontrolle der Individualvariablen) variiert.

wurde in einem nächsten Schritt um die verwendeten Kovariaten erweitert. Hier berichten wir Befunde, die sich auf die Effekte der Schülermerkmale Geschlecht und Migrationshintergrund sowie des Schulmerkmals Schulform beziehen.

Tabelle 1: Testheftdesign von PISA 2006

	Booklet												
	B01	**B02**	**B03**	**B04**	**B05**	**B06**	**B07**	**B08**	**B09**	**B10**	**B11**	**B12**	**B13**
Position 1	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**M1**	**M2**	**M3**	**M4**	**R1**	**R2**
Position 2	**S2**	**S3**	**S4**	**M3**	**S6**	**R2**	**R1**	**M2**	**S1**	**M4**	**S5**	**M1**	**S7**
Position 3	**S4**	**M3**	**M4**	**S5**	**S7**	**R1**	**M2**	**S2**	**S3**	**S6**	**R2**	**S1**	**M1**
Position 4	**S7**	**R1**	**M1**	**M2**	**S3**	**S4**	**M4**	**S6**	**R2**	**S1**	**S2**	**S5**	**M3**

Anmerkungen: S = Science (Naturwissenschaften), M = Mathematics (Mathematik), R = Reading (Lesen)

In die Auswertung gingen insgesamt N = 33.480 Schülerinnen und Schüler ein, wobei wir eine kombinierte Stichprobe herangezogen haben, die aus den Populationen der 15-jährigen Schülerinnen und Schüler und der Neuntklässlerinnen und Neuntklässler zusammengesetzt war. Um eine interpretierbare Abschätzung der Schulformunterschiede zu ermöglichen, haben wir uns auf Schülerinnen und Schüler der alten Bundesländer konzentriert (mit Ausnahme des Saarlandes) und zwischen den traditionellen Schulformen des dreigliedrigen Sekundarschulsystems (d. h. Hauptschule, Realschule und Gymnasium) sowie der zusammengefassten Gruppe der integrierten Gesamtschule und der Schulen mit mehreren Bildungsgängen unterschieden. Schülerinnen und Schüler mit spezifischem Förderbedarf wurden aus den Auswertungen ausgeschlossen.

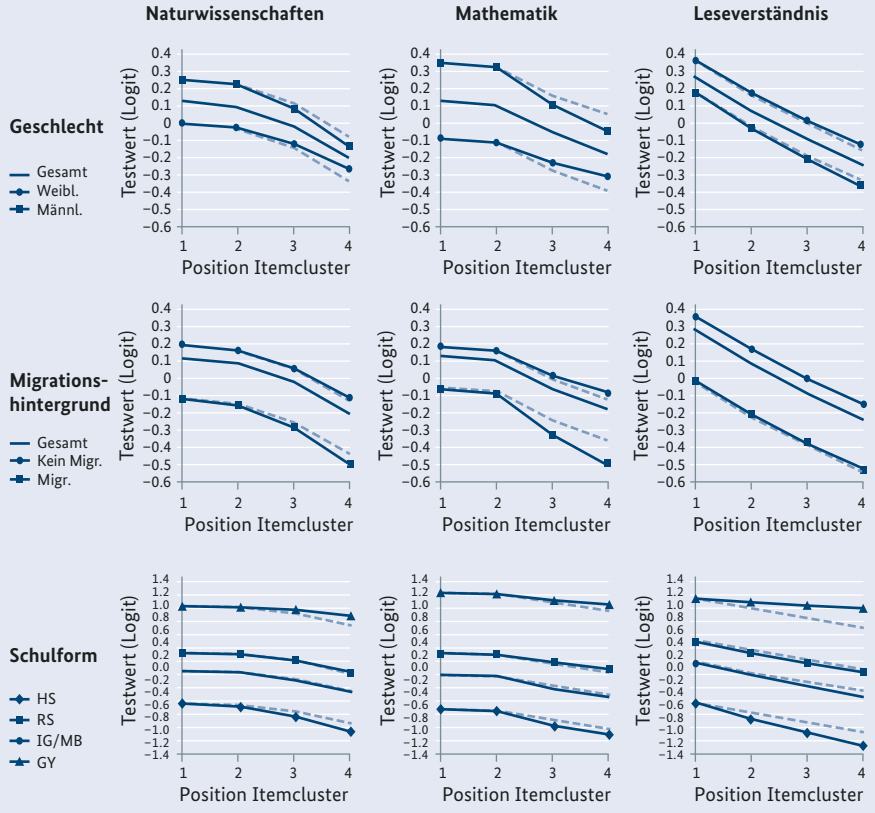
In Abbildung 3 sind die über die Itemcluster gemittelten Testleistungsverläufe in Abhängigkeit der Kovariaten Geschlecht und Migrationshintergrund sowie der Clusterposition dargestellt. Die grau dargestellten Linien beschreiben die mittleren Kompetenzverläufe in der Gesamtstichprobe. Die gestrichelten Linien geben den erwarteten Verlauf in jeder Gruppe unter der Annahme konstanter Positionseffekte wider. Abweichungen zwischen den gruppenspezifischen Kompetenzverläufen (schwarze durchgezogene Linien) und den gestrichelten Linien indizieren somit die Verzerrung, die sich aus der Ausblendung gruppenspezifischer Positionseffekte ergibt.

Aus Abbildung 3 geht hervor, dass sich die Gestalt der mittleren Positionseffekte (d. h. Kompetenzverläufe) zwischen den Domänen unterscheidet. Die Positionseffekte weisen im Bereich Lesen eine lineare Form auf, während sie in den Bereichen Naturwissenschaften und Mathematik erst in der zweiten Testhälfte deutlich zutage treten (ab Clusterposition 3). Hinzu kommt, dass sich die auf der Logit-Metrik ausgedrückte Höhe der Positionseffekte (Differenz zwischen der ersten und letzten Clusterposition) zwischen den Domänen unterscheidet. Die Positionseffekte fielen im Bereich Lesen am höchsten aus.

Abbildung 3 stellt die Positionseffekte in Abhängigkeit der Kovariaten dar. Jungen zeichneten sich im Vergleich zu Mädchen durch stärkere Abnahmen in allen

Kompetenzbereichen aus. Dieser Unterschied trat insbesondere für den Bereich Mathematik zutage, während er im Bereich Lesen vernachlässigbar erscheint. Stärkere Kompetenzrückgänge fanden sich zudem für Schülerinnen und Schüler mit Migrationshintergrund in den Bereichen Naturwissenschaften und Mathematik, wobei der Unterschied im Bereich Mathematik erneut am deutlichsten ausgeprägt war. Hinsichtlich der besuchten Schulform fanden sich für alle drei Domänen vergleichbare Befunde. Die Positionseffekte sind in den Gymnasien weitgehend vernachlässigbar und in den Hauptschulen am stärksten akzentuiert. Die Positionseffekte unterschieden sich im Bereich Lesen am deutlichsten zwischen den Schulformen.

Abb. 3: Testwerte nach Domänen (Spalten), Subgruppen (Zeilen) und Positionen der Itemcluster (x-Achse)



Die in Abbildung 3 dargestellten Zusammenhänge implizieren, dass die in Abhängigkeit der Hintergrundmerkmale ermittelten Kompetenzunterschiede von der Clusterposition abhängen. Die für die Bereiche Naturwissenschaften und Mathematik hinsichtlich der ersten Clusterposition ermittelten Vorteile der Jungen reduzieren

sich über den Verlauf der Testbearbeitung. Die Kompetenzunterschiede zwischen Schülerinnen und Schülern mit und ohne Migrationshintergrund nehmen hingegen zu (Naturwissenschaften und Mathematik), und ein vergleichbarer Befund findet sich für die Schulformunterschiede.

Tabelle 2: Mittelwertunterschiede (Logit-Metrik) für die erste (P1), letzte (P4) und gemittelte Itemposition (Mittel) sowie prozentuale Veränderung der Unterschiede relativ zur ersten Position

	Naturwissenschaft			Mathematik			Lesen		
	P1	Mittel	P4	P1	Mittel	P4	P1	Mittel	P4
*Geschlecht (**männlich** vs. weiblich)*									
Effekt	0.25	0.21	0.12	0.44	0.37	0.26	−0.18	−0.21	-0.24
% Veränderung		(−20%)	(−51%)		(−17%)	(−41%)		(16%)	(30%)
Migrationshintergrund (mit vs. ohne)									
Effekt	−0.31	−0.34	−0.39	−0.24	−0.31	−0.42	−0.18	−0.38	−0.37
% Veränderung		(10%)	(26%)		(33%)	(77%)		(−2%)	(−3%)
*Schulformenver**gleiche***									
Gy vs. IG/MB									
Effekt	1.03	1.11	1.22	1.32	1.38	1.47	1.00	1.22	1.43
% Veränderung		(7%)	(19%)		(5%)	(11%)		(22%)	(42%)
GY vs. RS									
Effekt	0.74	0.80	0.88	0.98	1.00	1.03	0.66	0.83	1.00
% Veränderung		(7%)	(19%)		(2%)	(5%)		(27%)	(52%)
GY vs. HS									
Effekt	1.55	1.67	1.86	1.87	1.95	2.07	1.66	1.94	2.20
% Veränderung		(8%)	(20%)		(5%)	(11%)		(17%)	(33%)
IG/MB vs. RS									
Effekt	−0.29	−0.31	−0.34	−0,34	−0.38	−0.44	−0.35	−0.39	−0.43
% Veränderung		(7%)	(17%)		(12%)	(28%)		(13%)	(25%)
IG/MB vs. HS									
Effekt	0.52	0.56	0.63	0.55	0.57	0.61	0.65	0.71	0.77
% Veränderung		(9%)	(22%)		(5%)	(11%)		(9%)	(18%)
RS vs. HS									
Effekt	0.81	0.87	0.97	0.89	0.96	1.04	1.00	1.11	1.20
% Veränderung		(8%)	(21%)		(7%)	(17%)		(11%)	(20%)

Anmerkungen: GY = Gymnasium, IG/MB = Integrierte Gesamtschule und Schulen mit mehreren Bildungsgängen, RS = Realschule, HS = Hauptschule

Aus einer praktischen Perspektive stellt sich somit die Frage nach dem Grad der Verzerrung der Effektstärken, die sich bei einer Ausblendung der Positionseffekte ergeben. In Tabelle 2 sind die Leistungsunterschiede an den Extrempositionen 1 und 4 sowie die über alle Positionen gemittelten Unterschiede abgetragen. Die Unterschiede an Position 1 können als näherungsweise frei von Positionseffekten betrachtet werden, während die Ergebnisse an Position 4 maximal vom Positionseffekt betroffen sind. Die gemittelten Unterschiede entsprechen näherungsweise den in Large-Scale-Assessments ermittelten Effekten. Tabelle 2 dokumentiert, dass sich die Kompetenzunterschiede zwischen den Positionen 1 und 4 zum Teil deutlich un-

terscheiden, während die gemittelten Unterschiede sich in vielen Fällen vom Betrag her nicht wesentlich von Ergebnissen an der ersten Clusterposition unterschieden. Größere Unterschiede finden sich jedoch für den Effekt des Migrationshintergrunds (Mathematik) und der Schulformunterschiede (Lesen).

4.2 Zusammenfassung

Gegenstand von Studie 2 ist die Untersuchung von Positionseffekten in einem prototypischen Large-Scale-Assessment. Die Ergebnisse dokumentieren, dass die Testbearbeitungspersistenz nicht nur zwischen Individuen, sondern auch zwischen Einzelschulen variiert. Ein weiterer wichtiger Befund ist, dass sich die in Large-Scale-Assessments häufig betrachteten Teilpopulationen von Schülerinnen und Schülern in ihrer Testbearbeitungspersistenz voneinander unterscheiden. Dieser Befund impliziert, dass die Validität von Kompetenzvergleichen unter Umständen gefährdet sein könnte, da die ermittelten Kompetenzunterschiede nicht ausschließlich auf die zugrunde liegenden Kompetenzausprägungen, sondern darüber hinaus von den gruppenspezifischen Ausprägungen der Testbearbeitungspersistenz abhängen. Für die in dieser Studie betrachteten Hintergrundvariablen zeigte sich jedoch, dass die Verzerrungen in vielen Fällen vergleichsweise gering ausfallen. Dieser Befund ist zwei Ursachen geschuldet. Erstens zeichnen sich die Positionseffekte in den Bereichen Naturwissenschaften und Mathematik durch einen nicht linearen Verlauf aus, wonach die erste Hälfte der Tests nur im geringen Maß von Positionseffekten betroffen ist. In der Konsequenz üben die zum Teil deutlichen Gruppenunterschiede in den Positionseffekten im Mittel (d. h. über alle Positionen hinweg) einen vergleichsweise schwachen Einfluss auf das Gesamtergebnis aus. Zweitens waren die hinsichtlich der ersten Itemclusterposition ermittelten Effekte (die per Definition nicht von Clusterpositionseffekten betroffen sind) bereits relativ stark ausgeprägt, sodass sich die über die nachfolgenden Testteile gemittelten Positionseffekte relativ dazu nur gering auf den Gesamtmittelwert auswirken.

An dieser Stelle mag sich der Eindruck einstellen, dass Positionseffekte eher vernachlässigbare Konsequenzen für die in Schulleistungsstudien erzielten Ergebnisse aufweisen. Jedoch sei darauf verwiesen, dass dies nicht unbedingt gelten muss. Erstens waren einige Vergleiche durchaus in substanziellem Umfang von Positionseffekten betroffen (z. B. Migrationshintergrund im Bereich Mathematik und Schulformvergleiche im Bereich Lesen), sodass die erzielten Ergebnisse nicht generell als robust gegenüber Positionseffekten gelten können. Zweitens sind viele in Schulleistungsstudien untersuchte Effekte weitaus schwächer ausgeprägt als die hier betrachteten Zusammenhänge (z. B. Effekte familiärer Interaktionsstile). Sofern diese Variablen ähnlich stark wie die hier untersuchten Merkmale mit dem Positionseffekt assoziiert sind, können sich größere Validitätsprobleme ergeben. Drittens gilt anzumerken, dass PISA auf ein rotiertes Testheftdesign mit balancierten Itemclusterpostionen setzt. In diesem Design sind die gemessenen Kompetenzen über die mittlere Position definiert. Im Gegensatz dazu greifen viele andere groß angelegte Schulleistungsstudien auf eine festgelegte Sequenz von domänenspezifischen Tests zurück (z. B. Na-

turwissenschaft an Position 1, Mathematik an Position 2 usw.). In solchen Designs sind die auf Grundlage der zum Ende der Sequenz dargebotenen Tests stärker vom Effekt der Testbearbeitungspersistenz betroffen. Dies führt beispielsweise dann zu erheblichen Problemen, wenn die Testergebnisse kriteriumsorientiert (z. B. bezüglich Kompetenzstufen) interpretiert werden sollen. In Abhängigkeit der Position, auf der eine Domäne vorgegeben wird, kommt man zu unterschiedlichen Schlüssen, welche Anteile der Population bestimmte Dinge wissen und können. Insgesamt ergibt sich somit die Einschätzung, dass die durch die Positionseffekte induzierten Verzerrungen in Abhängigkeit des verwendeten Testheftdesigns in verschärfter oder abgemilderter Form zutage treten können (Weirich, Hecht & Böhme, 2014).

5 Studie 3: Itempositionseffekte im Längsschnitt

Retestdesigns sind in der empirischen Bildungsforschung weit verbreitet. An solche Studien ist die Hoffnung geknüpft, dass sie die Erfassung individueller Lerngewinne und deren Korrelate erlauben. Bis heute ist aber faktisch nichts über die Rolle von Itempositionseffekten in Längsschnittstudien bekannt. Im Prinzip können Itempositionseffekte die Ergebnisse von Längsschnittstudien in verschiedener Weise beeinflussen. So wirken sich Zu- oder Abnahmen in der mittleren Ausprägung von Positionseffekten auf die Schätzung der mittleren Lerngewinne aus. Insofern die Änderungen in den mittleren Ausprägungen der Positionseffekte zwischen verschiedenen zu vergleichenden Teilpopulationen (z. B. Schulformen) unterscheiden, ist zudem die Validität von Rückschlüssen über Gruppenunterschiede in Lerngewinnen gefährdet.

Da bis dato keine empirischen Befunde zur zeitlichen Entwicklung von Itempositionsffekten vorliegen (für eine aktuelle Studie siehe aber Nagy et al., 2015), können kaum empirisch fundierte Erwartungen über die Veränderung der Testbearbeitungspersistenz in Längsschnittstudien formuliert werden. So erscheinen Abnahmen in der mittleren Ausprägung von Positionseffekten in vielen Schulleistungsstudien plausibel, da ältere Schülerinnen und Schüler über ein größeres Portfolio mentaler Ressourcen verfügen, die negativen Positionseffekten entgegenwirken können (z. B. Decodiergeschwindigkeit). Andererseits könnten sich negative Positionseffekte über die Zeit verstärken, da sich die wiederholte Teilnahme an einer Testung motivationshemmend auswirken könnte. Hinsichtlich der Reteststabilität von Positionseffekten erscheinen auch unterschiedliche Szenarien plausibel. So könnten diese eine relativ zeit- und situationsüberdauernde Größe repräsentieren oder sich durch eine hohe Situationsspezifität auszeichnen.

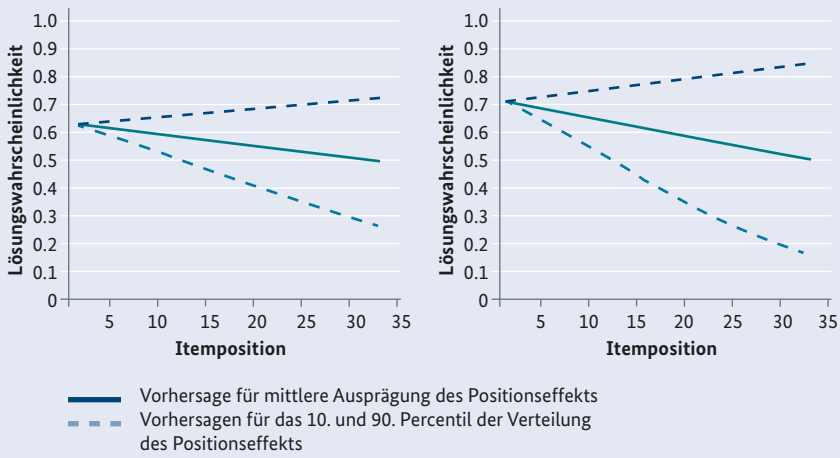
Die hier skizzierten Fragen stehen im Zentrum von Studie 3 (Nagy, Rose & Nagengast, 2015), die ebenso wie Studie 1 exemplarisch für den Bereich Leseverständnis durchgeführt wurde. Zu diesem Zweck wurde erneut auf die TRAIN-Daten zurückgegriffen (*N = 3.092),* wobei hier Leistungswerte der in den Klassenstufen 5 und 6 durchgeführten Erhebungen eingingen. In dieser Studie wurde die besuchte Schulform (Haupt-, Real- und Mittelschule) als Kovariate verwendet, da Schulformunterschiede in Lerngewinnen häufig im Fokus längsschnittlicher Schulleistungsstudien stehen.

5.1 Zentrale Ergebnisse

Die Auswertung der Daten geschah auf Grundlage einer Erweiterung des von Debeer und Janssen (2013) vorgeschlagenen IRT-Modells für Positionseffekte. Unsere Erweiterung vereinigt IRT-Modelle für wiederholte Messungen (von Davier, Xu & Carstensen, 2011) mit dem Modell für Positionseffekte. Das in Studie 1 verwendete Modell erlaubt zwar eine vollständigere Erfassung von Itempositionseffekten, seine Umsetzung für den Längsschnittfall erwies sich jedoch als problematisch, da es eine große Zahl nicht linearer Parameterrestriktionen beinhaltet.

Abbildung 4 beschreibt die messzeitpunktspezifischen Ausprägungen der Positionseffekte anhand der vorhergesagten Lösungswahrscheinlichkeiten für ein typisches Item (mittlere Schwierigkeit und Itemdiskrimination). Aus der Abbildung geht hervor, dass die mittlere Ausprägung der Itempositionseffekte in Klassenstufe 6 deutlich ansteigt. Ebenso wurde ein prägnanter Anstieg der Variabilität der Positionseffekte festgestellt. In diesem Modell zeigte sich, dass die individuellen Unterschiede in den Positionseffekten eine geringe Stabilität aufweisen (r = .22; SE = 0.11; p = .040). Dieser Befund indiziert, dass die individuelle Ausprägung der Itempositionseffekte eine hohe Situationsspezifität aufweist. Die hinsichtlich der ersten Itemposition definierte Traitvariable zeichnete sich hingegen durch ein hohe Stabilität aus (r = .74; SE = 0.03; p < .001).

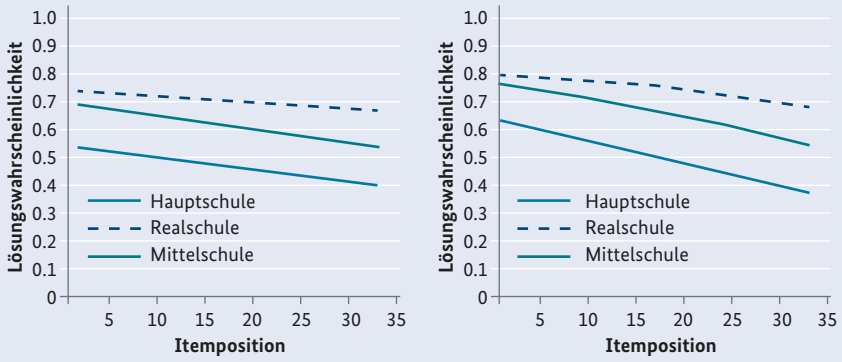
Abb. 4: Lösungswahrscheinlichkeit eines prototypischen Items in Abhängigkeit der Itemposition in den Klassenstufen 5 (linke Teilabbildung) und 6 (rechte Teilabbildung)



In Abbildung 5 werden die Schulformunterschiede anhand der Lösungswahrscheinlichkeit eines prototypischen Items für beide Messzeitpunkte dargestellt. Wie ersichtlich wird, unterschieden sich Leistungen zwischen den Schulformen zu beiden Messzeitpunkten deutlich voneinander. Ein interessanter Befund ist, dass die Leistungsunterschiede in Klassenstufe 5 nur im geringen Maß auf Unterschie-

de in Itempositionseffekten zurückzuführen sind (d. h. relativ parallel verlaufende Lösungswahrscheinlichkeiten über Positionen). Die Ergebnisse für den zweiten Messzeitpunkt unterscheiden sich davon, da die negativen Itempositionseffekte in der Gruppe der Schülerinnen und Schüler an Haupt- und Mittelschulen deutlicher ausgeprägt sind. In diesen Gruppen wurde ein markanter Anstieg der Itempositionseffekte ermittelt. Dieser Befund indiziert, dass die Ausblendung von Itempositionseffekten eventuell zu einer verfälschten Einschätzung über schulformspezifische Lerngewinne führen kann. So wurde der Lerngewinn in Hauptschulen ohne Berücksichtigung des Itempositionseffekts auf $d = 0.42$ (relativiert an der Ausgangsmessung) geschätzt. Die Effektstärke stieg nach Berücksichtigung des Positionseffekts um rund 25 Prozent auf $d = 0.52$ an.

Abb. 5: Lösungswahrscheinlichkeit eines Items mit mittlerer Schwierigkeit und Diskrimination in Abhängigkeit der Itemposition und der besuchten Schulform. Darstellungen für die Klassenstufen 5 (linke Teilabbildung) und 6 (rechte Teilabbildung)



5.2 Zusammenfassung

Gegenstand von Studie 3 ist die Untersuchung von Itempositionseffekten in einem längsschnittlichen Setting. Die Befunde sind für die angewandte Schulleistungsforschung von Bedeutung, da sie zeigen, dass die wiederholte Erfassung von Schülerinnen und Schülern zu Testartefakten führen kann. Der hier festgestellte Anstieg in den mittleren Itempositionseffekten kann sich, sofern er nicht berücksichtigt wird, in einer Unterschätzung der Lerngewinne niederschlagen. Ebenso können Zusammenhänge von Hintergrundmerkmalen mit den Itempositionseffekten zu einer falschen Einschätzung ihrer Effekte auf die Leistungsentwicklung führen.

Die in Studie 3 berichteten Befunde wurden in der Essenz kürzlich in einem unabhängigen Datensatz repliziert. Nagy und Kollegen (2015) konnten einen Anstieg der Positionseffekte in einer Large-Scale-Studie replizieren, wobei die Veränderung in nicht gymnasialen Schulformen besonders akzentuiert ausfiel und die Ausblendung von Positionseffekten sich deutlich auf die Abschätzung der Leistungszuwächse auswirkte.

An dieser Stelle kann festgehalten werden, dass die Auswirkungen von Positionseffekten in Längsschnittstudien besonders gravierend ausfallen können. Der Grund hierfür ist, dass Lerngewinne in späteren Phasen der Beschulung (z. B. in der gymnasialen Oberstufe) relativ gering ausfallen (z. B. Bloom, Hill, Black & Lipsey, 2008), sodass eine valide Interpretation der entsprechenden Effekte eine hohe Präzision der Zuwachsschätzung voraussetzt. Ebenso gilt, dass die Zusammenhänge der Zuwächse in den Testwerten mit anderen Kovariaten im Vergleich zu querschnittlichen Zusammenhängen gering sind, sodass Itempositionseffekte zu fehlerhaften Rückschlüssen über die Determinanten von Lerngewinnen führen könnten.

6 Abschließendes Resümee

Itempositionseffekte sind im Kontext von Low-Stakes-Large-Scale-Assessments eher die Regel als die Ausnahme. Derartige Effekte wurden bis vor kurzer Zeit statistisch als feste Effekte behandelt, die als eine Eigenschaft des verwendeten Tests und nicht der getesteten Schülerinnen und Schüler konzipiert sind. In Übereinstimmung mit der aktuellen Literatur (z. B. Debeer & Janssen, 2013) wurde im vorliegenden Projekt die Sichtweise eingenommen, dass Itempositionseffekte eher als ein auf der Personenseite lokalisiertes Phänomen zu behandeln sind, da sie im Sinne individueller Reaktionen auf Leistungstests zu verstehen sind. Ziel war es, den personenseitigen Aspekt von Itempositonseffekten genauer zu untersuchen. Konkret liefert das vorliegende Projekt Beiträge zur Erfassung von Itempositionseffekten, zu den Korrelaten von Itempositionseffekten, deren Auswirkungen auf die anhand herkömmlicher Testwerte vollzogenen Inferenzen und zur Rolle von Itempositionseffekten in Längsschnittstudien. Darüber hinaus umfasst das vorliegende Projekt eine weitere Teilstudie (Rose, Nagy, Nagengast, Frey & Becker, 2015), die sich den Interaktionen zwischen Positionseffekten und Auswirkungen der Abfolge von Testitems zu unterschiedlichen Inhaltsdomänen widmet (Brennan, 1992; Harris, 1991).

Die in diesem Beitrag zusammengefassten Studien liefern zusätzliche Evidenz für die Existenz individueller und schulischer Unterschiede in Positionseffekten. Die von uns durchgeführten Studien erweitern den bisherigen Kenntnisstand zu Positionseffekten in wichtiger Weise, da sie dokumentieren, dass diese systematisch mit anderen Personen- und Schulmerkmalen und Merkmalen der Testsituation assoziiert sind. Insgesamt unterstreichen unsere Befunde die Einschätzung, dass Itempositionseffekte als Indikatoren der Testbearbeitungspersistenz betrachtet werden können.

Die Erfassung der Korrelate der Testbearbeitungspersistenz ist nicht nur aus einer grundlagenwissenschaftlichen Perspektive interessant. Für anwendungsorientiert arbeitende Wissenschaftlerinnen und Wissenschaftler ist besonders wichtig, dass individuelle, gruppenspezifische und zeitliche Unterschiede in der Testbearbeitungspersistenz eine Gefährdung der Validität vieler Schlussfolgerungen darstellen. Auch wenn der Einfluss von Positionseffekten häufiger – aber eben nicht immer – vom Betrag her klein ausfällt, sind wir der Meinung, dass von Itempositionseffekten erhebliche Probleme bezüglich der Validität von Testwertinterpretationen bei Large-

Scale-Assessments ausgehen. Dies gilt insbesondere für Studien, die auf Testheftdesigns mit balancierten Itempositionen setzen, wie z. B. PISA. Unsere Ergebnisse deuten darauf hin, dass die Situation in Studien mit einem domänensequenziellen Design wahrscheinlich gravierender ausfällt (vgl. Tabelle 2). Nichtsdestotrotz plädieren wir dafür, dass die Auswirkungen von Positionseffekten auf die Ergebnisse von Leistungsvergleichen – wann immer es möglich ist – untersucht werden sollten.

7 Ausblick

Itempositionseffekte sind in allen aktuellen Arbeiten als graduelle Abnahmen der Lösungswahrscheinlichkeit von Items konzipiert (z. B. Debeer & Janssen, 2013). Statistisch gesehen werden Itempositionseffekte durch die Rate der Veränderung der Itemschwierigkeiten indiziert. Im vorliegenden Projekt sind wir diesem Ansatz gefolgt. Nichtsdestotrotz erscheinen neben der Rate der Veränderung auch andere Indikatoren der Testbearbeitungspersistenz sinnvoll. So könnte diese auch über die Position in einem Test, ab der eine Reduktion der Lösungswahrscheinlichkeit festzustellen ist, indiziert werden (z. B. Yamamoto & Gitomer, 1993). Der Vergleich unterschiedlicher Indikatoren der Testbearbeitungspersistenz ist ein wichtiges Thema, da andere Aspekte der Änderung des Lösungsverhaltens die vorliegenden Leistungsdaten unter Umständen besser als die Rate der Veränderung beschreiben und darüber hinaus höher mit individuellen Kovariaten korreliert sein könnten.

Ein weiterer Aspekt, der eine verstärkte Beachtung verdient, richtet sich auf die Gestaltung optimaler Testheftdesigns, die den Einfluss von Positionseffekten auf die Ergebnisse der Leistungsmessung minimieren und/oder die Identifikation und statistische Kontrolle von Itempositionseffekten optimieren (vgl. Hecht, Weirich, Siegle & Frey, 2015). Da eine nahezu vollständige Ausschaltung von Positionseffekten nur mithilfe vergleichsweise kurzer Tests zu bewerkstelligen ist und darüber hinaus nicht mit einer sequenziellen Darbietung von Tests zu unterschiedlichen Leistungsdomänen zu vereinen ist, ist eine designbedingte (approximativ) vollständige Kontrolle von Positionseffekten nicht möglich. Ein alternativer Ansatz könnte im Versuch bestehen, eine Balance zwischen von per Design relativ unbeeinflussten Testteilen und Testteilen, die eine statistische Kontrolle von Positionseffekten ermöglichen, zu erreichen.

Schließlich verdient die Betrachtung von Itempositionseffekten in Längsschnittsettings eine intensivere Beachtung. In den vergangenen Jahren ist auch im Bereich der Schulleistungsforschung eine verstärkte Zuwendung hin zu Längsschnittstudien festzustellen. An derartige Studien wird die Hoffnung geknüpft, dass sie einen detaillierten Einblick in Lernprozesse erlauben. Wie wir zeigen konnten, erscheinen längsschnittliche Schulleistungsstudien als anfällig für Positionseffekte. Die bisherigen Erfahrungen beschränkten sich aber auf einfache Retestdesigns mit zwei Erhebungspunkten. Zukünftige Forschung sollte die längerfristige Entwicklung der Testbearbeitungspersistenz untersuchen.

Literaturverzeichnis

Bloom, H. S., Hill, C. J., Black, A. R. & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research in Educational Effectiveness, 1*, 289–328.

Brennan, R. L. (1992). The Context of Context Effects. *Applied Measurement in Education, 5*, 225–264.

von Davier, M., Xu, X. & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76*, 318–336.

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

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

Frey, A., Carstensen, C. H., Walter, O., Rönnebeck, S. & Gomolka, J. (2008). Methodische Grundlagen des Ländervergleichs. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006 in Deutschland: Die Kompetenzen der Jugendlichen im dritten Ländervergleich* (S. 375–397). Münster: Waxmann.

Frey, A., Hartig, J. & Rupp, A. (2009). Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice, 28,* 39–53.

Harris, D. (1991). Effects of passage and item scrambling on equating relationships. *Applied Psychological Measurement, 15*, 247–256.

Hecht, M., Weirich, S., Siegle, T. & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement*. Advance online publication.

Johnson, E. G. (1990). *National assessment of educational progress: Design of the 1992 assessment*. Princeton, NJ: Educational Testing Service.

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

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.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161.

Nagengast, B., Nagy, G., Rose, N. & Becker, M. (2015). *Positionseffekte in den Leistungstests der nationalen Erweiterung der PISA 2006 Studie: Eine Mehrebenenstudie zu individuellen und kontextuellen Prädiktoren von Positionseffekten.* Vortrag auf der 12. Tagung der Fachgruppe Methoden & Evaluation der Deutschen Gesellschaft für Psychologie, Jena.

Nagy, G., Lüdtke, O., Köller, O., Heine, J.-H. & Mang. J. (2015). *IRT Skalierung der Leistungstests in der PISA-Längsschnittstudie 2012/2013: Konsequenzen von Positionseffekten auf die Abschätzung der Leistungsentwicklung.* Vortrag auf der 3. Tagung der Gesellschaft für Empirische Bildungsforschung, Bochum.

Nagy, G., Rose, N. & Nagengast, B. (2015). *Reteststabilität von Itempositionseffekten in einem Leseverständnistest: Mittelwertstabilität, und Entwicklung individueller Unterschiede von Klassenstufe 5 zu 6.* Vortrag auf der 12. Tagung der Fachgruppe Methoden & Evaluation der Deutschen Gesellschaft für Psychologie, Jena.

Nagy, G., Rose, N. & Trautwein, U. (2013). *Individuelle Unterschiede in der Testermüdung während der Bearbeitung eines Leseverständnistests: Eine Anwendung eines IRT-Modells zur Erfassung individueller Positionseffekte und ihrer Korrelate.* Vortrag auf der 1. Tagung der Gesellschaft für Empirische Bildungsforschung, Kiel.

Rose, N., Nagy, G., Nagengast, B., Frey, A. & Becker, M. (2015). *Multiple Itemkontexteffekte in Mehrdimensionalen IRT-Modellen: Modellierung von Effekten der Itemposition, der Blockposition und der Domänenabfolge.* Vortrag auf der 12. Tagung der Fachgruppe Methoden & Evaluation der Deutschen Gesellschaft für Psychologie, Jena.

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

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*, 114–128.

Yamamoto, K. & Gitomer, D. H. (1993). Application of a HYBRID model to a test of cognitive skill representation. In N. Fredriksen, R. Mislevy & I. Beijar (Hrsg.), *Test theory for a new generation of tests* (S. 275–296). Hillsdale, NJ: Erlbaum.