Hartig, Johannes; Jude, Nina; Wagner, Wolfgang

Methodische Grundlagen der Messung und Erklärung sprachlicher Kompetenzen

Klieme, Eckhard [Hrsg.]: Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie. Weinheim u.a. : Beltz 2008, S. 34-54

urn:nbn:de:0111-opus-31526

in Kooperation mit:

BELTZ

http://www.beltz.de

Johannes Hartig / Nina Jude / Wolfgang Wagner

3 Methodische Grundlagen der Messung und Erklärung sprachlicher Kompetenzen

Die Analyse der komplexen und umfangreichen Daten, wie sie aus Large-Scale-Assessments wie DESI resultieren, ist mit einer Reihe methodischer Herausforderungen verbunden. Verschiedene statistische Analyseverfahren, die dem aktuellen Stand der Forschungsmethodik entsprechen, kommen zum Einsatz, um eine möglichst gesicherte Interpretation der Ergebnisse zu gewährleisten. Das vorliegende Kapitel soll einen Überblick über die verschiedenen methodischen Anforderungen geben, die bei der Analyse der DESI-Daten zu berücksichtigen waren, und zugleich darüber informieren, mit welchen Verfahren gearbeitet wurde. Der erste Teil des Kapitels befasst sich mit der Messung der sprachlichen Kompetenzen, d.h. mit der Auswertung der Daten aus den Leistungstests. Hier werden kurz die verwendeten Skalierungsmodelle skizziert sowie die Schätzung der Kompetenzwerte und deren Aufbereitung für eine anschauliche Ergebnisdarstellung behandelt. Im zweiten Teil des Kapitels geht es um Fragen, die bei den weiteren Auswertungsschritten zu berücksichtigen waren, in erster Linie bei der Untersuchung von Zusammenhängen der Kompetenzen mit weiteren Schüler-, Unterrichts- und Schulvariablen. Das Kapitel kann die methodischen Grundlagen der verwendeten Analyseverfahren nicht detailliert darstellen, gibt jedoch eine Übersicht über die Verfahrenslogik der eingesetzten Techniken und verweist auf vertiefende Literatur. Methodisch vorgebildeten Leserinnen und Lesern werden Hinweise auf Verfahrensdetails gegeben, die für eine Replikation der Ergebnisse benötigt werden.

3.1 Auswertung der Leistungstests und Schätzung der Kompetenzen

Skalierungsmodelle

Die Schülerinnen und Schüler bearbeiteten in DESI eine Vielfalt verschiedener Aufgaben. Die Antworten auf diese Aufgaben erfolgten sowohl in geschlossener Form (Multiple Choice) als auch in freier schriftlicher Form. Diese Antworten wurden nach vorher definierten Kodierschemata auf ihre Korrektheit bzw. auf ihre Güte hin bewertet (vgl. Kapitel 2). Von der Güte der Testantworten der Schüler wird auf die Kompetenzen der Schüler in verschiedenen Bereichen des Deutschen und Englischen geschlossen. Je mehr Aufgaben eines Tests ein Schüler korrekt beantwortet, desto höher wird seine Kompetenz im jeweiligen Bereich eingeschätzt.

Die Kompetenztests in DESI wurden in einem Matrix-Design vorgegeben, bei dem jeder Schüler nur einen Teil der Aufgaben jedes Tests bearbeitet; zudem wurden von jedem Schüler zu Beginn und zum Ende der neunten Klasse nicht diesel-

ben, sondern unterschiedliche Aufgaben bearbeitet (vgl. Kapitel 2). Es ist daher nicht möglich, die bloße Anzahl gelöster Aufgaben in einem Bereich als einen Indikator für die jeweils interessierende Kompetenz zu verwenden, da dies die unterschiedlichen Schwierigkeiten der Aufgaben nicht berücksichtigen würde. Zur Schätzung der Schülerkompetenzen werden daher Modelle der Item-Response-Theorie (IRT) verwendet. In diesen Modellen werden die Schwierigkeiten der Testaufgaben und die Kompetenzen der getesteten Personen auf derselben Kompetenzskala beschrieben; zwischen der Kompetenz der Schüler und den erwarteten Lösungshäufigkeiten der jeweiligen Aufgaben werden statistische Zusammenhänge formuliert. Die Auswertung mit IRT-Modellen erlaubt eine gemeinsame Schätzung der Kompetenzen aller Schüler auf derselben Skala, auch wenn diese unterschiedliche Aufgaben bearbeitet haben (z.B. Rost 2004). Voraussetzung hierfür ist, dass die empirischen Daten sich mit dem verwendeten IRT-Modell hinreichend gut beschreiben lassen. Dies kann durch geeignete statistische Anpassungsmaße (*fit indices*) für einzelne Aufgaben oder ein komplettes Modell geprüft werden.

Das in DESI eingesetzte spezifische Skalierungsmodell ist ein generalisiertes Rasch-Modell (Adams/Wilson/Wang 1997; Adams/Wu 2007), welches in der Analysesoftware ConQuest implementiert ist (Wu/Adams/Wilson 1998). Innerhalb dieses Modells können sowohl dichotome (z.B. falsch/richtig) als auch ordinale Auswertungsformate (z.B. falsch/teilweise gelöst/vollständig gelöst) innerhalb desselben Tests modelliert werden. Das Raschmodell nimmt für alle Aufgaben gleich starke Zusammenhänge zwischen Kompetenz und Lösungswahrscheinlichkeit an. In Begriffen der klassischen Testtheorie bedeutet dies die Annahme gleicher Trennschärfen, in faktorenanalytischen Begriffen gleiche Faktorladungen aller Aufgaben. Diese Annahme verlangt zwar eine strengere Auswahl geeigneter Aufgaben als weniger restriktive IRT-Modelle, erlaubt aber eine konsistentere Beschreibung der Kompetenzskala unter Bezug auf die Aufgabenschwierigkeiten (Wilson 2003). Der angenommene Zusammenhang zwischen Kompetenz und Lösungswahrscheinlichkeit ist für eine zweistufige Antwortauswertung in Abbildung 3.1 graphisch veranschaulicht. Dabei ist die Schwierigkeit einer Aufgabe im Raschmodell als der Punkt auf der Kompetenzskala definiert, an dem die Aufgabe von Personen mit einer gleich hohen Kompetenz zu 50% gelöst wird. Umgekehrt können die individuellen Kompetenzen von Personen in Form von Lösungswahrscheinlichkeiten interpretiert werden: Personen, deren Kompetenz so hoch ist wie die Schwierigkeit einer Aufgabe, sollten diese Aufgabe, wenn das Modell gilt, zu 50% lösen.

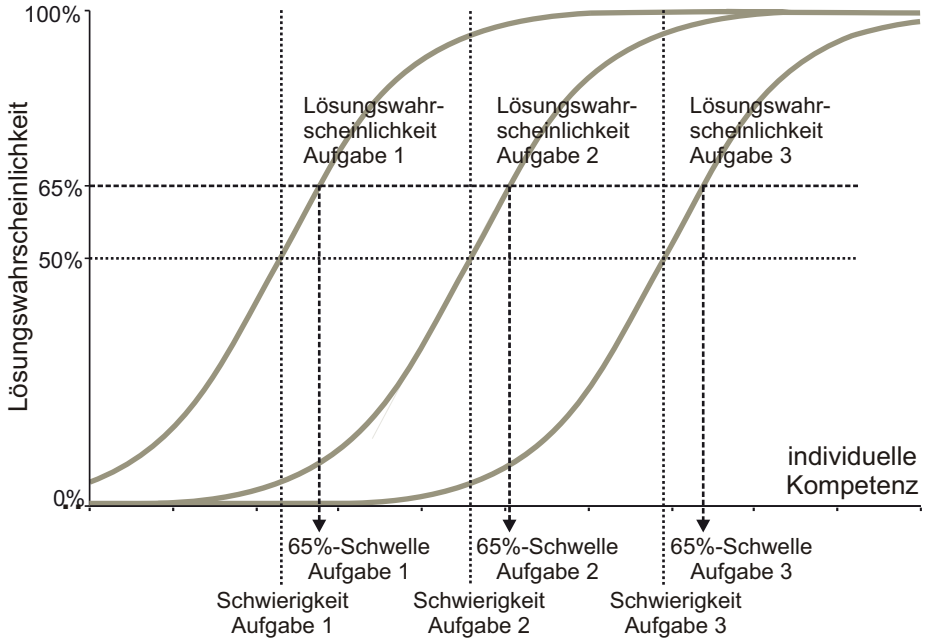

Abbildung 3.1: Veranschaulichung des im Raschmodell angenommenen Zusammenhangs zwischen individueller Kompetenz, Aufgabenschwierigkeit und Lösungswahrscheinlichkeit.

Die Beschreibung von Aufgabenschwierigkeiten und Kompetenzen auf derselben Skala und die Interpretierbarkeit in Form von Lösungswahrscheinlichkeiten stellt einen zentralen Vorteil der Auswertung mit IRT-Modellen dar und wird auch bei der Definition der Kompetenzniveaus (s.u.) herangezogen. Bei der Beschreibung der in DESI erfassten Kompetenzen interessiert, welche Schüler welche sprachlichen Anforderungen hinreichend sicher bewältigen können. Da eine Lösungshäufigkeit von 50% als Kriterium für das Beherrschen spezifischer Anforderungen zu niedrig ist, wurde in DESI eine Lösungswahrscheinlichkeit von 65% als Kriterium zur Verankerung der Aufgaben auf den Kompetenzskalen gewählt, wie sie z.B. auch in der Third International Mathematics and Science Study (TIMSS) verwendet wurde (vgl. Klieme/Baumert/Köller/Bos 2000)[1]. Die im Raschmodell beschriebene Beziehung zwischen Kompetenz, Aufgabenschwierigkeit und Lösungswahrscheinlichkeit erlaubt es, diese „65%-Schwellen" einfach zu ermitteln, indem zu der Aufgabenschwierigkeit ein konstanter Wert addiert wird. In Abbildung 3.1 ist die Bildung der 65%-Schwellen illustriert.

1 Zum Vergleich: In den PISA-Studien wird eine Lösungswahrscheinlichkeit von 62% als Kriterium für eine hinreichende Beherrschung der Aufgaben angesetzt, im US-amerikanischen National Assessment of Educational Progress (NAEP) sogar eine Lösungswahrscheinlichkeit von 80%.

Bildung von Messwerten für die erfassten Kompetenzen

Innerhalb der IRT stehen verschiedene Techniken zur Verfügung, mit der numerische Messwerte für die Kompetenzen der getesteten Personen ermittelt werden können. Eine zentrale Grundidee der verschiedenen Verfahren ist, dass die individuellen Kompetenzen so gebildet werden, dass die Wahrscheinlichkeit der tatsächlich beobachteten Antworten maximiert wird – dieses Prinzip wird als *Maximum-Likelihood-Schätzung (ML)* bezeichnet. Im Raschmodell sind die auf ML basierenden Kompetenzwerte in erster Linie eine Funktion der Anzahl gelöster Aufgaben, d.h. je mehr Aufgaben eine Person gelöst hat, desto höher ist ihre Kompetenz. Bei der Auswertung von Daten aus einem Matrixdesign spielt zusätzlich noch eine Rolle, welche Aufgabenmenge eine Person bearbeitet hatte, so dass berücksichtigt wird, ob die Menge gelöster Aufgaben auf leichteren oder schwierigeren Aufgaben basiert. Auf ML basierende Messwerte sind im Hinblick auf die Ermittlung individuell fairer Kompetenzwerte optimal, sind jedoch, wie alle individuellen Messwerte, mit einem unvermeidlichen Messfehler behaftet. In DESI wurden auf Basis der Testergebnisse als ML-basierte Kompetenzwerte *Weighted Likelihood Estimates* (*WLEs*; Warm 1989) geschätzt. Diese auf der ML-Methode basierenden Schätzer gewichten die Messwerte mit der jeweils individuell verfügbaren Testinformation, um so die messfehlerbedingte Verzerrung der Populationsverteilung auszugleichen. Die individuellen Messwerte sind jedoch ebenso messfehlerbehaftet wie ungewichtete ML-Schätzer.

Das Ziel von Large-Scale-Assessments wie DESI ist jedoch nicht die Messung von Kompetenzen der einzelnen getesteten Schüler, wie dies z.B. bei Abschlussprüfungen oder Vergleichsarbeiten der Fall ist. Es geht vielmehr darum, die Verteilungen der Kompetenzen in der interessierenden Grundgesamtheit (d.h. aller Neuntklässler) und in interessierenden Untergruppen (z.B. differenziert nach Geschlecht oder Bildungsgang) zu beschreiben. Diese beiden Zielsetzungen – Individualdiagnostik und Populationsbeschreibung – stehen insofern in einem gewissen Konflikt, als individuelle Messwerte für die erfassten Kompetenzen immer mit einem Messfehler behaftet sind. Dies führt dazu, das ML-basierte Messwerte zwar für die individuelle Kompetenzbeschreibung optimal sind, nicht jedoch, um die Verteilung der Kompetenzen und Zusammenhänge mit weiteren Personenmerkmalen in der interessierenden Grundgesamtheit unverzerrt zu beschreiben (Mislevy/Beaton/Kaplan/Sheehan 1992). Um dem Ziel von DESI als repräsentativer Studie gerecht zu werden und die Verteilung sprachlicher Kompetenzen in der Gesamtheit der Neuntklässler sowie die Zusammenhänge mit Schüler- und Unterrichtsmerkmalen optimal zu beschreiben, wird in DESI, wie auch in anderen Large-Scale-Assessments, von der Plausible-Value-Technik Gebrauch gemacht (Mislevy/Beaton/Kaplan/Sheehan 1992; Adams/Wu/Carstensen 2007). Hierbei werden bei der Schätzung der Kompetenzen nicht nur die Antworten auf die Testaufgaben herangezogen, sondern auch Eigenschaften der untersuchten Personen wie z.B. Geschlecht, Erstsprache oder Bildungsgang. Noch vor der Bestimmung individueller Kompetenzwerte werden die Zusammenhänge

zwischen diesen *Hintergrundvariablen* und der zu messenden Kompetenz ermittelt, z.B. ob Mädchen durchschnittlich höhere Werte haben als Jungen. Diese Zusammenhänge, die nicht auf der Schätzung individueller Werte basieren, werden im so genannten *Hintergrundmodell* auf latenter Ebene und damit messfehlerfrei geschätzt. Als Hintergrundvariablen wurden in DESI alle Variablen aus allen eingesetzten Fragebögen (Schüler, Eltern, Lehrer, Schulleiter und Fachkollegien; vgl. Kapitel 2) sowie die individuellen Leistungen im Kognitiven Fähigkeitstest (KFT, Heller/ Perleth 2000) einbezogen. Zusätzlich wurde, um der für Schulleistungsdaten charakteristischen Stichprobenstruktur gerecht zu werden (s.u.), der mittlere Anteil gelöster Aufgaben in einer Klasse im jeweils interessierenden Test in das Hintergrundmodell einbezogen. Um die große Menge dieser sowohl auf individueller Ebene als auch auf Klassen- und Schulebene erfassten Daten in das Hintergrundmodell einbeziehen zu können, wurden die Daten mittels Hauptkomponentenanalysen zu unkorrelierten Faktorwerten zusammengefasst, die insgesamt 99.9% der Variation in den ursprünglichen Variablen erfassten. Das so konstruierte Hintergrundmodell berücksichtigt praktisch alle linearen Zusammenhänge, die zwischen den in DESI erfassten Kompetenzen und anderen Variablen bestehen. Zusätzlich wurden für die Faktoren Bildungsgang, Geschlecht und Erstsprache auch die Wechselwirkungen erster und zweiter Ordnung aufgenommen.

Anschließend wurden für die untersuchten Schülerinnen und Schüler *Plausible Values (PVs)* erzeugt; zusätzlich wurden auch die jeweiligen WLEs für die erfassten Kompetenzen ermittelt. Die PVs berücksichtigen sowohl die Menge gelöster Aufgaben als auch die Eigenschaften der untersuchten Personen. Wenn im ersten Schritt z.B. ein Kompetenzvorsprung für Mädchen ermittelt wurde, wird dieser bei der Schätzung der PVs dahingehend berücksichtigt, dass Mädchen etwas höhere Werte zugewiesen bekommen als Jungen mit der gleichen Anzahl gelöster Aufgaben. Dieses Vorgehen ist offensichtlich nicht zur Schätzung fairer individueller Messwerte konzipiert, erlaubt jedoch eine unverzerrte Schätzung der Kompetenzverteilungen in der zugrunde liegenden Population sowie der Zusammenhänge zwischen individuellen Hintergrundmerkmalen und der untersuchten Kompetenz. Es ist ein Charakteristikum der Plausible-Value-Technik, die auf der multiplen Imputation fehlender Werte basiert (Schafer 1997), dass für jede interessierende Kompetenz mehrere – im Fall von DESI fünf – PVs generiert werden (vgl. auch Walter 2006). Analysen auf Basis der PVs werden mit jedem PV wiederholt, und die Streuung zwischen den fünf Analyseergebnissen wird bei der Einschätzung der statistischen Bedeutsamkeit der Ergebnisse berücksichtigt.

Im DESI-Datensatz lagen also für jede sprachliche Kompetenz zwei Typen von Kompetenzschätzungen vor, nämlich WLEs und PVs. Je nach Art der Analyse kommen beide Werte zur Anwendung. Die weitaus meisten Analysen, bei denen die Zusammenhänge zwischen den Schülerkompetenzen und weiteren individuellen, Unterrichts- oder Schulvariablen von Interesse sind, werden mit den PVs durchgeführt, da diese die besten Schätzungen für die Populationsverteilungen der Kompetenzen liefern. Eine simultane Erzeugung von PVs für alle Kompetenzen in einer gemeinsa-

men mehrdimensionalen Analyse war in DESI durch die große Anzahl verschiedener Kompetenzen technisch nicht möglich. Für die Untersuchung von Zusammenhängen zwischen den verschiedenen sprachlichen Kompetenzen werden daher die jeweiligen WLEs verwendet, da die PVs aus separaten Skalierungen stammen und damit die Zusammenhänge zwischen den verschiedenen Kompetenzen nicht korrekt wiedergeben. Die mittels WLEs geschätzten Zusammenhänge sind zwar messfehlerbehaftet, dies kann jedoch durch die Modellierung mit latenten Variablen berücksichtigt werden (vgl. Kapitel 18; Hartig/Jude 2005). Auch die Rückmeldungen an die Lehrkräfte der in DESI untersuchten Schulklassen (vgl. auch Kapitel 4) wurden auf Basis der WLEs vorgenommen.

Für alle in DESI eingesetzten Tests wurden jeweils separat individuelle Werte in Form von WLEs und PVs gebildet. Zusätzlich wurde auf Basis der Tests, die zum Ende der neunten Jahrgangsstufe eingesetzt worden waren, jeweils ein Gesamtwert für Deutsch und Englisch gebildet. Hierzu wurden die Daten der einzelnen Tests in einer gemeinsamen Skalierung – d.h. wie Antworten auf einen einzigen großen Test – analysiert. Die einzige Ausnahme stellte hierbei der C-Test zur Textrekonstruktion dar. Diese Lückentexte waren zur Bildung der Kompetenzwerte auf Ebene der einzelnen Lücken skaliert worden (Harsch/Schröder 2007). Da der Textrekonstruktions-Test hierdurch sehr viele „Aufgaben" enthält, wäre er bei der Bildung des Gesamttestwertes gegenüber den übrigen, kürzeren Tests übergewichtet worden. Um dies zu vermeiden und dennoch die wesentlichen Informationen über Leistungsunterschiede in Textrekonstruktion im Englisch-Gesamtwert zu berücksichtigen, wurde die Leistung aus den einzelnen Lückentexten zu Kategorien (0 bis 5 richtig gefüllte Lücken, 6 bis 10 richtig gefüllte Lücken, usw.) zusammengefasst und die einzelnen Texte als mehrstufige Items behandelt. Sowohl für Deutsch als auch für Englisch wurden als Schätzung für die Gesamtkompetenz der Schüler in der jeweiligen Sprache WLEs und PVs geschätzt. Diese Gesamtwerte für Deutsch und Englisch werden dort als abhängige Variablen verwendet, wo eine Differenzierung nach den einzelnen Kompetenzen aus inhaltlicher Sicht nicht angezeigt ist und mit einer Verwendung der Gesamtwerte eine übersichtlichere Ergebnislage erreicht werden kann. Beispiele sind die Effekte von Unterrichts- und Schuleigenschaften (vgl. Kapitel 26 bis 34).

Zu Beginn der neunten Jahrgangsstufe wurde lediglich eine Auswahl der DESI-Tests eingesetzt. Es ist daher nicht möglich, für diesen Zeitpunkt einen Gesamttestwert zu bilden, der mit der Gesamtleistung am Ende der neunten Jahrgangsstufe vergleichbar ist; aus diesem Grund wurden auch keine PVs für eine Gesamtleistung zu Beginn der neunten Jahrgangsstufe gebildet. Für einige Analysen ist es jedoch erstrebenswert, das Leistungsniveau der Schülerinnen und Schüler zu Beginn des Untersuchungszeitraums als Kontrollvariable zu verwenden. Um dies zu ermöglichen, wurde als Indikator für das Ausgangsniveau der sprachlichen Kompetenzen zu Beginn der neunten Jahrgangsstufe für Deutsch und Englisch jeweils der Mittelwert der WLEs für die zu Beginn der Untersuchung eingesetzten Tests gebildet.

Normierung der Kompetenzskalen

Die Schätzungen für individuelle Kompetenzwerte werden in Analysen mit dem in DESI verwendeten IRT-Modell auf einer so genannten Logit-Skala dargestellt. Diese Werte liegen typischerweise in einem numerisch relativ kleinen Wertebereich um null (z.B. -3 bis 3), die genaue Streuung hängt von den Messeigenschaften des jeweiligen Tests ab. Um anschaulichere Werte für die in DESI gemessenen Kompetenzen zu erhalten, wurden die für die weiteren Analysen und die deskriptiven Ergebnisdarstellungen verwendeten Werte so normiert, dass der Mittelwert der Grundgesamtheit *am Ende der neunten Jahrgangsstufe* auf 500 Punkte und die Standardabweichung auf 100 Punkte gesetzt wurde. Ein Schüler mit einem Wert von 550 Punkten für eine bestimmte Kompetenz würde in diesem Bereich also eine halbe Standardabweichung über dem Mittelwert aller Neuntklässler liegen. Die Skalierung mit einem Mittelwert von 500 und einer Standardabweichung von 100 Punkten ist an die Normierung der Kompetenzwerte in den PISA-Studien angelehnt (z.B. OECD 2001, 2004). Es ist jedoch zu beachten, dass in PISA der internationale Mittelwert der teilnehmenden OECD-Länder als Referenz verwendet wurde, und der Mittelwert der deutschen Schüler z.B. in Lesen in PISA 2000 ca. 480 Punkte betrug. In DESI, wo kein internationaler Vergleich vorgenommen wird, entsprechen 500 Punkte dem Mittelwert der deutschen Neuntklässler. Die numerischen Werte auf Basis dieser Normierung sind nicht mit den Werten aus den PISA-Studien vergleichbar. Die Verteilung am Ende der neunten Jahrgangsstufe wurde als Referenz für die Normierung gewählt, da zum Ende der neunten Jahrgangsstufe fast alle Tests eingesetzt wurden, zu Beginn der neunten Jahrgangsstufe nur eine Auswahl (vgl. auch Kapitel 2). Die Normierung der Messwerte zu Beginn der neunten Jahrgangsstufe wurde ebenfalls auf Basis der Verteilungen zum Ende der neunten Jahrgangsstufe vorgenommen. Dies führt dazu, dass sich für die zu Beginn der neunten Jahrgangsstufe gemessenen Kompetenzen normierte Mittelwerte kleiner als 500 ergeben, wenn innerhalb der neunten Jahrgangsstufe ein Zuwachs stattgefunden hat. Eine Ausnahme von diesem Vorgehen bildet der Test für Englisch Sprachbewusstheit im Bereich Soziopragmatik (Nold/Rossa 2007), der nur zu Beginn der neunten Jahrgangsstufe eingesetzt worden war. Die Messwerte für diesen Test wurden auf Basis der Werteverteilung zu Beginn der neunten Jahrgangsstufe normiert.

Die Normierung der Kompetenzskalen erfolgte auch für die Itemschwierigkeiten. Die 65%-Schwellen (s.o.) für alle Items wurden in der gleichen Weise transformiert wie die Messwerte für die jeweiligen Tests. Aufgabenschwierigkeiten und Kompetenzen werden so weiterhin auf derselben Skala beschrieben. Die Transformation der 65%-Schwellen führt dazu, dass zum Beispiel eine Aufgabe mit einer transformierten Schwierigkeit von 500 Punkten von Schülerinnen und Schülern mit einer durchschnittlichen Kompetenz zu ca. 65% gelöst wird. Eine Aufgabe mit einer transformierten Schwierigkeit von 600 Punkten wird nur von Schülerinnen und Schülern, deren Kompetenz eine Standardabweichung über dem Mittelwert liegt, mit einer Häufigkeit von 65% gelöst. In den Kapiteln zu den Deutsch- und

Englischkompetenzen (Kapitel 5-15) werden, soweit auf Aufgabenschwierigkeiten Bezug genommen wird, immer diese transformierten Aufgabenschwierigkeiten verwendet.

Schätzung von Kompetenzzuwächsen

In DESI wurden Schülerkompetenzen in ausgewählten Bereichen zu Beginn und am Ende der neunten Jahrgangsstufe erfasst. Ein zentrales Anliegen der Studie ist es, die Zuwächse, die in den verschiedenen sprachlichen Kompetenzen zwischen diesen Zeitpunkten erzielt werden, einzuschätzen sowie mögliche Prädiktoren für Unterschiede in den Zuwächsen zu untersuchen. In der IRT gibt es verschiedene Techniken, mit Daten aus mehreren Messzeitpunkten umzugehen (Rost 2004; Meiser/Stern/Langeheine 1998). Wenn Schätzungen der Kompetenzen zu mehreren Zeitpunkten miteinander verglichen werden sollen, ist es wichtig, dass die Schwierigkeiten der Aufgaben zu den verschiedenen Zeitpunkten im Modell gleich gesetzt werden. Eine Technik, dies bei der Schätzung individueller Kompetenzwerte für mehrere Messzeitpunkte zu erreichen, ist die Bildung sogenannter *virtueller Personen*. Die Messungen zum zweiten und weiteren Zeitpunkten werden hierbei so behandelt, als ob zusätzliche Personen den Test bearbeitet hätten (vgl. Rost 2004). Während diese Methode für ML-basierte Kompetenzschätzungen geeignet ist, führt sie bei der Schätzung von PVs zu stark verzerrten Verteilungsschätzungen, da die Abhängigkeiten zwischen den Antworten derselben Personen nicht berücksichtigt werden (Hartig/Kühnbach 2006). Diesen Abhängigkeiten wird Rechnung getragen, wenn die Daten aus den verschiedenen Messzeitpunkten in einem mehrdimensionalen Modell analysiert werden, in dem jeder Zeitpunkt als eine separate Kompetenzdimension behandelt wird. Durch ein Gleichsetzen der Aufgabenschwierigkeiten zwischen den Zeitpunkten kann erreicht werden, dass die Kompetenzschätzungen zwischen den Zeitpunkten vergleichbar sind. Werden auf Basis eines solchen mehrdimensionalen Modells PVs generiert, ermöglichen diese im Unterschied zu ML-basierten Kompetenzwerten eine messfehlerbereinigte Schätzung der Zusammenhänge sowohl zwischen den beiden Zeitpunkten als auch zwischen anderen Variablen und dem Kompetenzzuwachs zwischen den Zeitpunkten (Hartig/Kühnbach 2006; Hartig/Frey 2006).

In DESI wurden für die Tests, die zu beiden Messzeitpunkten in der neunten Jahrgangsstufe eingesetzt wurden, zunächst in einer Skalierung mit virtuellen Personen, d.h. in einem eindimensionalen Modell unter Einbezug aller vorhandenen Daten, die Aufgabenschwierigkeiten geschätzt. In einem zweiten Schritt wurde unter Verwendung dieser Aufgabenschwierigkeiten ein zweidimensionales Modell geschätzt, in dem die Kompetenz zu Beginn der neunten Jahrgangsstufe als eine, die Kompetenz am Ende der neunten Jahrgangsstufe als zweite latente Dimension behandelt wurde. Das Hintergrundmodell wurde wie oben beschrieben konstruiert. Die auf Basis dieses zweidimensionalen Modells generierten PVs wurden als Kompetenzschätzungen für die beiden Messzeitpunkte in der neunten

Jahrgangsstufe verwendet. Zusätzlich wurden die Differenzen zwischen den jeweils korrespondierenden PVs der beiden Zeitpunkte gebildet. Diese PV-Differenzen dienen als Schätzungen für den Kompetenzzuwachs zwischen Anfang und Ende des Schuljahres.

Definition der Kompetenzniveaus

Die für DESI vorgenommenen Kompetenzschätzungen sind quantitative Werte auf kontinuierlichen Skalen, die eine sehr feine Abstufung von Kompetenzunterschieden erlauben. Bei der Untersuchung von Zusammenhängen der Kompetenzen mit anderen Variablen ist diese Differenzierung von großem Vorteil. Es ist hingegen schwer, die Bedeutung unterschiedlich hoher Kompetenzwerte bezogen auf die konkreten inhaltlichen Anforderungen der Testaufgaben in dieser Differenziertheit anschaulich zu beschreiben (Beaton/Allen 1992). Für die Interpretation der DESI-Ergebnisse ist es jedoch sehr wichtig, welche konkreten Anforderungen im sprachlichen Bereich die untersuchten Schüler bewältigen können – es soll eine kriterienorientierte Interpretation der beobachteten Testleistung ermöglicht werden (z.B. Goldhammer/Hartig 2007; Rauch/Hartig 2007). Um eine solche anschauliche Interpretation der gemessenen Kompetenzen zu erreichen, werden die kontinuierlichen Kompetenzskalen in Abschnitte unterteilt (z.B. Beaton/Allen 1992, Klieme/Baumert/Köller/Bos 2000, OECD 2001, 2004). Diese Abschnitte werden als Kompetenzniveaus oder Kompetenzstufen[2] bezeichnet. Für jedes Kompetenzniveau wird dann eine anschauliche Beschreibung der Anforderungen vorgenommen, welche die Schüler, deren Leistungen im entsprechenden Abschnitt der Skala liegen, bewältigen können. Entscheidend bei der Definition der Kompetenzniveaus ist, an welchen Stellen der Skala die Schwellen zwischen den Niveaus gesetzt werden.

In DESI wurden zur Setzung der Schwellen zwischen den Kompetenzniveaus systematisch definierte Merkmale der Testaufgaben verwendet. Diese Aufgabenmerkmale, die für jeden der einzelnen Tests im vornherein definiert wurden, beschreiben spezifische Anforderungen, hinsichtlich derer sich leichtere von schwereren Aufgaben unterscheiden sollten. Die theoretisch angenommenen Zusammenhänge der Aufgabenmerkmale mit den tatsächlich ermittelten Aufgabenschwierigkeiten können in Regressionsanalysen untersucht werden. Aus diesen Analysen lassen sich auch erwartete Schwierigkeiten für bestimmte Kombinationen von Aufgabenmerkmalen ableiten. Diese erwarteten Schwierigkeiten wurden in DESI herangezogen, um die Schwellen zwischen den Kompetenzniveaus zu definieren. Hierdurch ist es möglich, die gemessenen Kompetenzen bezogen auf die sprachlichen Anforderungen

2 Angesichts möglicherweise irreführender Konnotationen des Stufenbegriffs (vgl. Helmke/Hosenfeld 2004) wurde für DESI der Begriff des Kompetenzniveaus dem der Kompetenzstufe vorgezogen. Die Kompetenzniveaus in DESI bezeichnen jedoch dasselbe wie Kompetenzstufen im Kontext anderer Studien, nämlich Abschnitte auf kontinuierlichen Kompetenzskalen, die mit dem Ziel einer kriteriumsorientierten Beschreibung der erfassten Kompetenzen gebildet werden.

in den verschiedenen Kompetenzbereichen zu beschreiben. Das generelle Vorgehen zur Definition der Kompetenzniveaus in DESI ist ausführlich bei Hartig (2007) beschrieben, die Aufgabenmerkmale für die einzelnen Kompetenzen in den jeweiligen Kapiteln in Beck/Klieme (2007).

3.2 Analyse von statistischen Zusammenhängen

Umgang mit fehlenden Werten

Es ist für Large-Scale-Assessments wie DESI typisch, dass die erfassten Daten fehlende Werte aufweisen. Zum Teil werden Daten aus ökonomischen Gründen von vornherein nur unvollständig erhoben (*missing by design*), wie z.B. die im Matrix-Design vorgegebenen Testaufgaben oder die fachspezifischen Fragen in den Schülerfragebögen. Weitere Daten fehlen aufgrund unvollständiger Beteiligungen an den Schüler- und Elternbefragungen (zu Erhebungsdesign und Beteiligungsraten s. Kapitel 2). Fehlende Werte aufgrund unterschiedlicher Beteiligungsraten fehlen in der Regel nicht zufällig, sondern hängen systematisch mit bestimmten Merkmalen von Schülern und Eltern zusammen; so sind z.B. die Beteiligungsraten an der Elternbefragung bei Eltern mit nicht deutscher Erstsprache geringer als bei deutschsprachigen Eltern[3]. Bei Analysen mit derartigen systematisch fehlenden Werten würde ein Ausschluss der Fälle mit fehlenden Werten aus der Analyse zu einer Verzerrung der Analyseergebnisse führen (z.B. Wolf 2006). Um unverzerrte Ergebnisse zu erhalten, können die fehlenden Werte mit einer *Multiplen Imputation* geschätzt werden (Schafer 1997). Hierbei werden vorhandene Informationen über Zusammenhänge zwischen den interessierenden Variablen sowie vorhandene Informationen über die Fälle mit teilweise fehlenden Werten zur Schätzung dieser fehlenden Werte herangezogen.

Für die Sprachkompetenzen erfolgte eine solche Schätzung der fehlenden Werte bei der oben beschriebenen Erzeugung der PVs, da die Plausible-Values-Technik auf der Multiplen Imputation basiert. Jeder Schüler und jede Schülerin, für den / die zu einem der beiden Messzeitpunkte in der neunten Jahrgangsstufe ein bearbeiteter Test vorlag, wurde als Teilnehmer/-in an der Studie definiert. Für diese $N = 10543$ Fälle wurden auf Basis des oben beschriebenen Hintergrundmodells PVs für die erfassten Kompetenzen generiert, auch wenn der jeweils interessierende Test nicht bearbeitet wurde. Auf diese Weise wurde vor allem die Datenbasis für die Schätzung der Kompetenzzuwächse erweitert, da PVs für beide Zeitpunkte auch für Fälle generiert wurden, die nur zu einem der beiden Zeitpunkte an der Testung teilgenommen hatten. Der Anteil der durch die Bildung der PVs geschätzten fehlenden Werte beträgt für alle Tests unter 10%. Dieser Anteil ist vergleichsweise gering, selbst bei ca. 40% fehlenden Werten in einzelnen Variablen kann eine Multiple Imputation noch zu bes-

3 Eine Regression mit den Variablen Erstsprache (Deutsch vs. nicht deutsch), Migrationsstatus, höchster Bildungsabschluss der Eltern und der Deutschleistung (inklusive aller Interaktionsterme) erklärt etwa 8% der Varianz im Fehlen des Elternfragebogens.

seren Analyseergebnissen führen als ein Ausschluss der Fälle aus der Analyse (z.B. Schafer/Olsen 1998).

Auch für eine zentrale Hintergrundvariable in DESI, den höchsten sozioökonomischen Status in der Familie (*Highest International Socio-Economic Index*, HISEI) nach Ganzeboom u.a. (1992), wurde eine Schätzung fehlender Werte vorgenommen. Dieser Index basiert auf Elternangaben. Für etwa 62% der Schülerinnen und Schüler lagen die dafür erforderlichen Angaben vor. Für weitere 29%, für die ein beantworteter Schülerfragebogen vorlag, wurde der Index geschätzt. Die Imputation erfolgte mit dem Programm *IVEWare* (Raghunathan/Solenberger/Van Hoewyk 2002). Auch hier wurden – analog zu den Plausible Values – fünf Werte generiert. Dazu wurde neben verschiedenen Schülerangaben, die in der gesamten Stichprobe erhoben wurden, insbesondere auf Angaben zum Beruf bzw. Schulabschluss der Eltern zurückgegriffen. Zur Berücksichtigung der Mehrebenenstruktur der Daten wurden die auf Klassenebene aggregierten Variablen (Klassenmittelwerte) als Faktorwerte (Hauptkomponentenanalyse, 24 Faktoren mit etwa 90% Varianzaufklärung) in die Imputation einbezogen.

Eine polytome Regression – inklusive einiger wichtiger Interaktionsterme – auf der Basis der Schülerangaben zum Beruf und dem Schulabschluss der Eltern erwies sich als optimales Verfahren bezüglich der Varianzaufklärung des HISEI. Nachteil dieses Verfahrens ist der verglichen mit der linearen Regression extrem hohe Rechenaufwand. Deshalb wurden fehlende Werte in den übrigen Variablen mittels linearer Regression imputiert. Erst in einem zweiten Schritt wurde dann der HISEI auf der Basis eines Datensatzes, bei dem – mit Ausnahme der Angaben zum Beruf bzw. Schulabschluss – alle Variablen bereits vollständig imputiert waren, mit Hilfe einer polytomen Regression imputiert.

Das gesamte Verfahren wurde anhand der vorliegenden Daten zum HISEI überprüft, indem analog zu den fehlenden Elternangaben Daten „gelöscht" und anschließend imputiert wurden. Es zeigte sich, dass die Ergebnisse von Zwei-Ebenen-Analysen auf Basis der „echten" bzw. der imputierten Daten sowohl bezüglich der Varianzen als auch hinsichtlich der Signifikanztests auf beiden Ebenen weitgehend vergleichbar sind. Zusätzlich wurden ausgewählte Zusammenhänge auf der Basis der gesamten Daten (mit fehlenden vs. mit imputierten HISEI-Werten) überprüft[4]. Auch hier zeigten sich vergleichbare Ergebnisse.

Bei allen übrigen Variablen und Skalen wurde auf eine solche aufwändige Imputation verzichtet. Einige Skalen aus den Schülerfragebögen wurden zwar nur bei der Hälfte der Schülerinnen und Schüler der jeweiligen Klassen erhoben, so dass auch dort eine Imputation auf den ersten Blick nahe gelegen hätte, dies betrifft allerdings nur die Zahl der fehlenden Werte. Da es sich hier um designbedingte „rein zufällig fehlende Werte" (*missing completely at random*, MCAR) handelt, sind die Schätzer für Mittelwerte, Standardabweichungen und Zusammenhänge ohnehin un-

4 Dazu wurden mit Hilfe des Programms Mplus 3.12 Korrelationsmatrizen auf zwei Ebenen mit der so genannten *missing at random maximum likelihood missing data*-Methode (MAR ML) geschätzt.

verzerrt (z.B. Wolf 2006). Bei Zusammenhangsanalysen wurde stattdessen auf das in der Analysesoftware Mplus implementierte Verfahren für Analysen mit fehlenden Werten zurückgegriffen.

Berücksichtigung der Stichprobenstruktur bei Signifikanztests und Standardfehlern

Ein wesentliches Ziel von DESI ist es, die Sprachkompetenzen in verschiedenen Gruppen von Schülern zu beschreiben, z.B. in verschiedenen Bildungsgängen oder mit verschiedenen sprachlichen Hintergründen. Wenn Unterschiede zwischen Gruppen zu beobachten sind, stellt sich die Frage, ob diese Unterschiede in der untersuchten Stichprobe nur zufällig zustande gekommen sind, oder ob sie auf die Gesamtheit der Neuntklässler übertragbar sind. Diese Frage wird mit statistischen Signifikanztests beantwortet. Hierbei wird geprüft, mit welcher Wahrscheinlichkeit ein vorgefundenes Ergebnis aus einer Population stammt, in welcher der interessierende Zusammenhang *nicht* vorliegt, d.h. die so genannte Nullhypothese gilt. Wenn diese Wahrscheinlichkeit unter einem vorher definierten Signifikanzniveau (üblicherweise 5%) liegt, wird die Nullhypothese verworfen und es wird davon ausgegangen, dass der Zusammenhang auch in der Grundgesamtheit, aus der die Stichprobe gezogen wurde, von null verschieden ist. Ein in diesem Zusammenhang wichtiges Konzept ist der *Standardfehler* statistischer Kennwerte. Dieser gibt die Genauigkeit an, mit der ein Wert – z.B. ein Mittelwert oder ein Korrelationskoeffizient – auf Basis der Stichprobendaten geschätzt werden kann. Je kleiner der Standardfehler, desto zuverlässiger kann von den Stichprobendaten auf die Population geschlossen werden. Mit Hilfe des Standardfehlers kann ein *Konfidenzintervall* definiert werden, das einen Bereich von ungefähr ± zwei Standardfehlern unter- und oberhalb des Mittelwertes umfasst und beschreibt, in welchem Wertebereich wahrscheinlich der interessierende Wert in der untersuchten Grundgesamtheit liegt. Je größer der Standardfehler und je breiter dadurch das Konfidenzintervall, desto ungenauer ist die Schätzung des jeweils interessierenden Kennwertes.

Bei der Schätzung von Standardfehlern und Signifikanztests werden in den Sozialwissenschaften in der Regel bestimmte Grundannahmen über das Verhältnis von Stichprobe und Grundgesamtheit gemacht. Die meisten häufig verwendeten Signifikanztests setzen voraus, dass die Fälle in der Stichprobe zufällig aus der Grundgesamtheit gezogen wurden. Diese Voraussetzung ist in DESI und vielen anderen Schulleistungsstudien nicht gegeben. Die interessierende Grundgesamtheit sind alle Schülerinnen und Schüler innerhalb Deutschlands in der neunten Klasse, gezogen wurden jedoch zunächst Schulen und dann Schulklassen innerhalb der Schulen (vgl. Kapitel 2). Innerhalb der Schulklassen wurden dann jeweils alle Schüler untersucht. Die resultierende Stichprobe hat eine so genannte *Klumpen-* oder *Clusterstruktur*, da die Fälle in Gruppen vorliegen, die sich untereinander ähnlicher sind (nämlich Schüler derselben Klassen) als zwischen den Gruppen (d.h. Schüler verschiedener Klassen). Wenn Daten mit einer solchen Clusterstruktur mit statisti-

schen Standardverfahren analysiert werden, führt dies zu einer Unterschätzung der Standardfehler, und Zusammenhänge werden zu schnell als signifikant betrachtet.

Es gibt zwei gängige Methoden, mit der Clusterstruktur von Stichproben angemessen umzugehen. Zum einen können die Standardfehler und Signifikanzen durch so genannte *Resampling*-Verfahren unter Berücksichtigung der Stichprobenstruktur ermittelt werden, wobei auch der Stichprobenplan und die Gewichtung einbezogen werden. Die statistischen Verfahren sind hierbei einfache Standardanalysen (z.B. Deskriptive Statistiken, Mittelwertsvergleiche, Korrelationsanalysen). Die entsprechenden Analysen werden jedoch mit systematisch variierender Gewichtung wiederholt, und die Variation der Ergebnisse zwischen den wiederholten Analysen wird in die Standardfehler und Signifikanztests mit einbezogen (Brick/Morganstein/Valliant 2000). Für DESI wurde zur Schätzung der Standardfehler und Signifikanzen bei einfachen Zusammenhangsanalysen die Software WesVar (Westat 2000) verwendet, die dieses leistet. Als Resampling-Methode wurde die *Balanced Repeated Replication* (BRR) nach Fay (1989) mit einem Perturbationsfaktor von 70% (K = 0.3; vgl. Westat 2000) verwendet. Diese Analysesoftware erlaubt neben der Berücksichtigung der Stichprobenstruktur auch simultan die Berücksichtigung der Streuungen zwischen den Plausible Values (s.o.).

Eine zweite Möglichkeit der Berücksichtigung der Clusterstruktur ist die Anwendung von *Mehrebenenmodellen* (z.B. Hox 2003; auch *Hierarchisch Lineare Modelle*). Hierbei wird der Umstand, dass die untersuchten Schüler in kompletten Klassen rekrutiert wurden, dadurch berücksichtigt, dass die Schulklasse als eine separate Analyseebene in die statistischen Modelle mit einbezogen wird. Die Unterschiede in den gemessenen Kompetenzen werden hierbei in Varianz *zwischen Klassen* („Ebene 2“) und in Varianz *zwischen Schülern* innerhalb von Klassen („Ebene 1“) aufgeteilt. Zusätzlich kann berücksichtigt werden, dass sich statistische Zusammenhänge zwischen verschiedenen Klassen unterscheiden können. Der Einsatz von Mehrebenenmodellen ist insbesondere dann angezeigt, wenn Variablen auf Klassenebene, wie z.B. Unterrichtseigenschaften, im Mittelpunkt des Interesses stehen. In DESI kamen Mehrebenenmodelle vor allem bei Analysen zu Unterrichtsmerkmalen und Schuleigenschaften zum Einsatz. Bei den Analysen zu Schuleigenschaften stellt die Schule die zweite Ebene dar, bei allen anderen Mehrebenenanalysen die Klasse. Eine theoretisch mögliche Analyse in Modellen mit drei Ebenen – Schüler in Klassen in Schulen – ist in DESI aufgrund der geringen Zahl von Schulklassen (maximal zwei) innerhalb der Schulen in der Regel wenig sinnvoll. Modelle mit drei Ebenen wurden daher nur zur Bestimmung der Varianzanteile auf Schul-, Klassen- und Individualebene (vgl. Kapitel 34), nicht jedoch bei Zusammenhangsanalysen verwendet. Die Mehrebenenanalysen in DESI wurden mit dem Programm Mplus (Muthén/Muthén 1998-2007) in der Version 4.0 durchgeführt. Einzelne Analysen, insbesondere die Schätzung von Varianzverteilungen zwischen drei Ebenen (Schüler, Klasse, Schule), wurden mit dem Programm HLM in der Version 6.02 vorgenommen. Diese Software erlaubt die Analyse von Plausible Values und Daten aus Multiplen Imputationen und die Berücksichtigung dieser Datenlage

bei der Schätzung von Standardfehlern und Signifikanzen. Die Berücksichtigung von Gewichten zum Ausgleich unterschiedlicher Teilnahmewahrscheinlichkeiten der Schülerinnen und Schüler (vgl. Kapitel 2) ist mit jedem der verwendeten Analyseprogramme möglich. Soweit nicht explizit anders angegeben, wurden alle in diesem Band berichteten Analysen unter Verwendung der in Kapitel 2 beschriebenen Gewichtung vorgenommen.

Während die Clusterstruktur der Stichprobe bei Analysen von Schülerdaten durch geeignete Methoden berücksichtigt werden muss, besteht dieses Problem bei Daten auf Schulebene nicht und auf Klassenebene in vernachlässigbarem Umfang. Die Schulen, an denen die Daten für DESI erhoben wurden, sind eine echte Zufallsstichprobe (vgl. Kapitel 2), und mit je zwei Klassen pro Schule ist die Clusterung der Stichprobe auf Klassenebene sehr gering. Bei Analysen auf Schul- und Klassenebene (z.B. ausschließlich mit Lehrer- oder Unterrichtsvariablen) kamen daher bei Signifikanztests auch Standardverfahren zum Einsatz.

Bestimmung von Effektgrößen

Die statistische Signifikanz eines Zusammenhangs oder Gruppenunterschieds bedeutet noch nicht zwingend, dass ein gefundener Effekt auch praktisch relevant ist, da bei großen Stichproben auch sehr schwache Effekte signifikant werden können. Zur Frage nach der Signifikanz eines Effektes kommt daher die Frage nach der *Effektgröße* hinzu. Zur Bestimmung der Effektgröße existieren für verschiedene Verfahren der Zusammenhangsanalyse verschiedene Kennwerte. Allen ist gemeinsam, dass untersuchte Effekte unabhängig von den ursprünglichen Skalen der untersuchten Variablen und in über verschiedene Analysen hinweg vergleichbaren Maßeinheiten beschrieben werden sollen. Ebenfalls gemeinsam ist den verschiedenen Effektgrößemaßen das Prinzip, einen untersuchten Zusammenhang oder Gruppenunterschied an der Streuung der interessierenden abhängigen Variablen zu relativieren. Die Größe eines Effektes wird also im Verhältnis zu den in der Stichprobe anzutreffenden Unterschieden zwischen den untersuchten Fällen interpretiert. Dieses Vorgehen lässt sich leicht an einem einfachen fiktiven Beispiel veranschaulichen. Wenn in einer Stichprobe ein Unterschied von zehn Testwert-Punkten zwischen Mädchen und Jungen gefunden wird, ist diese Zahl für sich genommen noch nicht aussagekräftig. Die Einschätzung der praktischen Bedeutsamkeit dieses Geschlechtereffekts, hängt von der Streuung der Testwerte ab. Wenn die Testwerte beispielsweise eine Streuung von SD = 250 haben, sind zehn Punkte eine vernachlässigbar kleine Differenz, nämlich nur 4% einer Standardabweichung. Hätten die Testwerte jedoch eine Streuung von z.B. nur 20 Punkten, wären zehn Punkte immerhin eine halbe Standardabweichung – die Geschlechtsdifferenz wäre in diesem Fall deutlich bedeutsamer.

Um die Größe von Effekten sowohl über verschiedene Studien als auch über verschiedene Analysemethoden hinweg vergleichen zu können, wird bei der Darstellung der Ergebnisse auf die übliche von Cohen (1988, 1992) vorgeschlagene

Klassifikation in kleine, mittlere und große Effekte zurückgegriffen. In Tabelle 3.1 sind die in den verschiedenen Ergebnisberichten verwendeten Effektgrößen sowie ihre Klassifikationen aufgelistet (vgl. Cohen 1988, 1992).

Tabelle 3.1: Definition und Klassifikation der bei der Ergebnisdarstellung verwendeten Effektgrößen.

			Klassifikation		
Analyse	**Effektgröße**	**Definition**	**klein**	**mittel**	**groß**
Gruppenunterschiede	d	An der Streuung standardisierte Differenz $d = \frac{M_A - M_B}{\sigma}$	0.20	0.50	0.80
Bivariate Zusammenhänge	r	An den Streuungen relativierte Kovarianz $r = \frac{\mathrm{cov}(x,y)}{\sigma(x)\cdot\sigma(y)}$	0.10	0.30	0.50
Multiple Regression	R^2	An der Gesamtvarianz relativierte durch alle Prädiktoren erklärte Varianz $R^2 = \frac{QS_{Modell}}{QS_{Gesamt}}$	0.02	0.13	0.26
Varianzanalysen	partielles η^2	An der Fehlervarianz relativierte Quadratsumme des Effekts $\eta^2 = \frac{QS_{Effekt}}{QS_{Effekt} + QS_{Fehler}}$	0.01	0.06	0.14

Anmerkung: QS = Quadratsumme

In Mehrebenenmodellen ist der Anteil der erklärten Varianz nicht so einfach zu bestimmen und zu interpretieren wie bei Analysen ohne Berücksichtigung der Mehrebenenstruktur (vgl. Hox 2003; Snijders/Bosker 1994). Bei den in diesem Band berichteten Ergebnissen aus Mehrebenenanalysen werden daher keine Effektgrößen angegeben, sondern für die einzelnen Effekte innerhalb der Modelle standardisierte Koeffizienten berichtet. Pfad- oder Regressionskoeffizienten werden hierfür an den Streuungen der jeweiligen unabhängigen und abhängigen Variablen standardisiert (vgl. Hox 2003), so dass sie in einem Wertebereich von -1 bis 1 liegen und von der Größenordnung entsprechend Korrelationskoeffizienten interpretiert werden können.

Modelle mit latenten Variablen

Bei der empirischen Erhebung von Schüler-, Klassen- und Schulmerkmalen werden häufig mehrere Indikatoren erfasst, welche auf dasselbe theoretische Konstrukt abzielen. Ein geläufiges Beispiel hierfür sind Fragebogenskalen, in denen dasselbe

Konstrukt (z.B. Leistungsmotivation) mit mehreren Fragen erfasst wird. Die Messung eines theoretischen Konstrukts mit Hilfe mehrerer verschiedener Indikatoren ist eine weit verbreitete Methode und dient vor allem einer höheren Messgenauigkeit (Reliabilität) und einer breiteren Generalisierbarkeit (Validität) der resultierenden Werte. Oft werden diese Messwerte durch Summen- oder Mittelwertsbildung aus den einzelnen Indikatoren gebildet, dieses Vorgehen wurde in DESI für die Fragebogendaten gewählt.

Es ist jedoch auch möglich, die Idee eines theoretischen Konstrukts, das mehreren beobachteten Variablen zugrunde liegt, direkt in statistische Analysemodelle einzubeziehen. In solchen Modellen werden die theoretischen Konstrukte als *latente Variablen* betrachtet, auf die über die gemessenen Indikatoren geschlossen werden kann. Modelle mit latenten Variablen erlauben eine messfehlerbereinigte Schätzung von Zusammenhängen zwischen den interessierenden Konstrukten. Auch die zur Analyse der Leistungstests verwendeten IRT-Modelle (s.o.) sind Modelle mit latenten Variablen, wobei die einzelnen Testaufgaben die beobachteten Indikatoren darstellen. Soweit in den Analysen zu DESI mit PVs gerechnet wird, werden die Kompetenzen statistisch also schon als latente Variablen behandelt. In vielen Analysen auf Basis von Fragebogendaten ist es ebenfalls so, dass mehrere beobachtete Variablen auf ein übergeordnetes Konstrukt zurückgeführt werden, z.B. verschiedene konkrete Formen der Kooperation in Fachkollegien auf die allgemeine Kooperativität an den jeweiligen Schulen (vgl. Kapitel 34). In diesen Fällen wurden Modelle mit latenten Variablen angewandt, in denen beobachtete numerische Werte (z.B. aus Fragebogenskalen) als Indikatoren für latente Variablen verwendet und die Zusammenhänge der interessierenden Konstrukte auf latenter Ebene untersucht wurden. Diese Modelle mit quantitativen Indikatoren werden in der Literatur häufig unter dem Begriff *Strukturgleichungsmodelle* beschrieben (z.B. Bollen 1989). Wenn die latenten Variablen nicht kontinuierlich, sondern kategorial sind, wird von *latenten Klassenanalysen* (z.B. Rost 2004) gesprochen. Die beobachteten Variablen werden in diesem Fall als Indikatoren für die Zugehörigkeit zu bestimmten Gruppen (latenten Klassen) betrachtet, wobei die Gruppenzugehörigkeit eine nicht latente – also nicht direkt beobachtbare – Variable darstellt.

Auch die Analysen mit latenten Variablen wurden in DESI mit Mplus 4.0 durchgeführt. Es ist möglich, Modelle mit latenten Variablen mit Mehrebenenmodellen zur Berücksichtigung der Stichprobenstruktur (s.o.) zu kombinieren, hieraus resultieren z.B. Mehrebenen-Strukturgleichungsmodelle, die in den Analysen zu Unterrichts- und Schulvariablen zum Einsatz kamen.

Bei der Analyse von Modellen mit latenten Variablen ist es wichtig zu beachten, wie gut das verwendete Modell mit den tatsächlich beobachteten Daten übereinstimmt. Um die Güte dieser Übereinstimmung zu beurteilen, können verschiedene statistische Indizes für die Modellanpassung (*model fit*) herangezogen werden. Diese Gütemaße basieren alle auf der Übereinstimmung der beobachteten Daten mit dem Datenmuster, dass bei Gültigkeit des geschätzten Modells zu erwarten wäre. Bei einigen dieser Gütemaße wie dem *Goodness of Fit Index* (GFI) sind möglichst hohe

Werte nahe eins wünschenswert, bei anderen, wie dem *Root Mean Square Error of Approximation* (RMSEA) möglichst niedrige Werte nahe Null. Tabelle 3.2 gibt einen Überblick darüber, welche Werte für verschiedene in diesem Band verwendete Gütemaße als Hinweise auf eine gute Modellanpassung interpretiert werden. Eine detaillierte Beschreibung verschiedener Gütemaße findet sich bei Schermelleh-Engel und Moosbrugger (2002).

Tabelle 3.2: Beurteilung der Güte von Modellen mit latenten Variablen anhand ausgewählter Gütekriterien.

Gütekriterium	Akzeptable Modellanpassung	Gute Modellanpassung
χ^2 / df (Chi-Quadrat-Prüfstatistik geteilt durch Freiheitsgrade)	≤ 3.00	≤ 2.00
SRMR (Standardized Root Mean Square Residual)	≤ 0.08	≤ 0.05
RMSEA (Root Mean Square Error of Approximation)	≤ 0.08	≤ 0.05
GFI (Goodness of Fit Index)	≥ 0.90	≥ 0.95
NNFI (Non-Normed Fit Index)	≥ 0.90	≥ 0.95
CFI (Comparative Fit Index)	≥ 0.90	≥ 0.95

Verwendung von Kontrollvariablen

Bei der Untersuchung der Zusammenhänge der Kompetenzen mit anderen Variablen ist zu berücksichtigen, dass die in DESI erhobenen Größen in vielfältiger Weise miteinander korreliert sind. Wenn Unterschiede in den sprachlichen Kompetenzen mit einer bestimmten Variablen erklärt werden sollen, muss immer berücksichtigt werden, dass diese Variable mit anderen möglichen erklärenden Größen *konfundiert* ist. Wenn z.B. der sozioökonomische Status der Eltern mit den kognitiven Grundfähigkeiten der Kinder zusammenhängt, muss dieser Abhängigkeit Rechnung getragen werden, wenn der spezifische Effekt des sozioökonomischen Status auf Unterschiede in sprachlichen Kompetenzen beurteilt werden soll. Hierzu kann der Effekt unterschiedlicher kognitiver Fähigkeiten statistisch kontrolliert werden, indem diese Variable z.B. in einer Regressionsanalyse zusammen mit dem sozioökonomischen Status als Prädiktor verwendet wird. In diesem Fall würde die kognitive Grundfähigkeit als *Kontrollvariable* verwendet. Der Effekt des sozioökonomischen Status würde dann als der Effekt interpretiert, den diese Variable hätte, wenn alle Schüler hinsichtlich ihrer kognitiven Grundfähigkeit gleich wären.

Welche Variable in einem Modell eine „Kontrollvariable" darstellt, ist eine jeweils aufgrund inhaltlicher Überlegungen zu entscheidende Frage. Der Einbezug von Kontrollvariablen ist in DESI insbesondere dort wichtig, wo schul- und unterrichtsbezogene Variablen mit Schülermerkmalen konfundiert sind, auf welche die Schule keinen Einfluss hat. Die Idee dieser Kontrolle lässt sich anschaulich ausdrükken, indem man z.B. positive Effekte bestimmter Unterrichtstechniken dahingehend

beschreibt, dass diese Praktiken mit *erwartungswidrig* guten Kompetenzen einhergehen. Damit ist gemeint, dass die Kompetenzen der Schüler in Klassen, in denen diese Unterrichtspraxis verstärkt zum Einsatz kam, höher sind, als z.B. aufgrund ihrer kognitiven Grundfähigkeiten und ihres Elternhauses zu erwarten gewesen wäre. Diese Abweichungen von den aufgrund der individuellen Ausgangsvoraussetzungen zu erwartenden Kompetenzen werden in der Schulforschung auch als *value added* bezeichnet, d.h. als der zusätzliche Beitrag zur Kompetenzentwicklung, der über außerhalb der Schule liegende Einflussfaktoren (z.B. die soziale Herkunft der Schülerinnen und Schüler) hinaus geht.

Neben den Schülermerkmalen werden in vielen Analysen zu DESI auch Unterschiede zwischen den verschiedenen Bildungsgängen kontrolliert. Dies ist wichtig, wenn z.B. bestimmte Unterrichtsmerkmale in Hauptschulen häufiger zu beobachten sind, und die Wirkung dieser Merkmale unabhängig davon eingeschätzt werden soll, dass in Hauptschulklassen insgesamt niedrigere Kompetenzen zu beobachten sind. Auch bei den auf individueller Ebene erfassten Kontrollvariablen kann die Variation auf verschiedenen Analyseebenen berücksichtigt werden. Variablen wie die kognitiven Grundfähigkeiten oder der soziale Status können einerseits auf individueller Ebene berücksichtigt werden, andererseits aber auch auf Schul- oder Klassenebene. Auf höherer Ebene wird dann die Zusammensetzung der Schüler in einer Klasse oder Schule kontrolliert, die entsprechenden Kontrollvariablen können hierzu durch Aggregation der individuellen Werte gewonnen werden.

Zu den in den Analysen der DESI-Daten verwendeten Kontrollvariablen gehören, soweit nicht anders angegeben, durchweg der Bildungsgang, der individuelle Sprachhintergrund (Erstsprache) und der Anteil der Schüler mit nicht deutscher Erstsprache, die kognitiven Grundfähigkeiten des einzelnen Schülers und der einzelnen Schülerin sowie der entsprechende Klassendurchschnitt, das Geschlecht sowie der Mädchenanteil in der Klasse, der sozioökonomische Status der Herkunftsfamilie und die soziale Komposition der Klasse. Der Bildungsgang wurde bei der Verwendung als Kontrollvariable in der Regel in zwei *Dummy-Variablen* kodiert, in denen zum einen die Zugehörigkeit zum Gymnasium, zum anderen die Zugehörigkeit zur Realschule mit eins, die jeweils anderen Bildungsgänge mit null kodiert wurden. Hierdurch wurden Hauptschule und IGS gemeinsam als Referenzkategorie verwendet.

Es ist zu beachten, dass die in DESI erhobenen Daten auch unter Berücksichtigung von vielfältigen Kontrollvariablen in praktisch keinem Fall eine sichere *kausale* Aussage darüber zulassen, auf welche Variablen beobachtete Kompetenzunterschiede ursächlich zurückgeführt werden können. Dennoch ist die Verwendung von Kontrollvariablen wichtig, um den wechselseitigen Abhängigkeiten und relativen Bedeutsamkeiten der verschiedenen erklärenden Variablen Rechnung zu tragen.

Vorhersage von Kompetenzzuwächsen

Durch die in DESI vorgenommene Messung zu zwei Zeitpunkten kann untersucht werden, welche Variablen mit Unterschieden in Kompetenzzuwächsen innerhalb

des neunten Schuljahres zusammenhängen. Diese Variablen können sowohl individueller Natur sein (gibt es bestimmte förderliche Bedingungen im Elternhaus), aber auch auf Klassenebene zu suchen sein (welche Unterrichtsmerkmale gehen mit einem hohen Kompetenzzuwachs einher) oder auf Schulebene liegen (welche Charakteristika der Fachkollegien gehen mit höheren Kompetenzzuwächsen einher). Die Vorhersage von Veränderungswerten – d.h. von Differenzen zwischen zwei Zeitpunkten – ist in empirischen Untersuchungen generell mit Schwierigkeiten verbunden, da diese Werte höhere Messfehler beinhalten und, bedingt durch Messfehler und durch zeitpunktspezifische Variation, häufig negativ mit dem Ausgangswert korreliert sind (z.B. Rost 2004). Angesichts dieser Problematik ist die Analyse von Veränderungswerten ein in der sozialwissenschaftlichen Forschung kontrovers diskutiertes Thema (z.B. Campbell/Kenny 2003; Allison 1990). Aufgrund der genannten problematischen Eigenschaften von Veränderungswerten werden diese häufig nicht als abhängige Variablen verwendet. Stattdessen werden die Werte zum Ende der neunten Jahrgangsstufe als abhängige Variable und die Werte zu Beginn der neunten Jahrgangsstufe als Prädiktor in das Modell einbezogen, um interindividuelle Unterschiede in den Ausgangswerten statistisch zu kontrollieren. Dieses sogenannte kovarianzanalytische Vorgehen führt jedoch unter bestimmten Bedingungen, insbesondere wenn der Ausgangswert mit einem interessierenden Prädiktor korreliert ist, zur Schätzung artifizieller Effekte – es resultieren Zusammenhänge von Drittvariablen mit den „kontrollierten" Messwerten zum zweiten Zeitpunkt, obwohl in Wahrheit keine Zusammenhänge mit den Zuwächsen zwischen den Zeitpunkten bestehen. Die Verwendung von Differenzwerten als abhängige Variable führt unter den meisten Bedingungen hingegen zu korrekten Schätzungen der interessierenden Effekte (Allison 1990; Hartig/Jude 2006, Jude/Hartig 2006). In DESI wurden daher, wenn der Kompetenzzuwachs von Interesse war, die PV-Differenzen für die zu beiden Messzeitpunkten in der neunten Jahrgangsstufe eingesetzten Tests als abhängige Variablen verwendet.

Ausnahmen bilden jene Analysen, in denen beide Messzeitpunkte in der neunten Jahrgangsstufe gezielt aufgenommen wurden, um zum Beispiel Effekte des Unterrichts in Abhängigkeit von der Leistung zum ersten Messzeitpunkt zu berechnen. In Mehrebenen-Pfadmodellen wurde in diesem Fall die Leistung zu Beginn und zum Ende der neunten Jahrgangsstufe in ihrem Zusammenhang mit Unterrichtsvariablen modelliert, um die Adaptivität des Unterrichts zu untersuchen. In diesen Modellen wurden für jeden der beiden Messzeitpunkte in der neunten Jahrgangsstufe Zusammenhänge zwischen der Leistung und den oben angeführten Hintergrundvariablen modelliert.

Literatur

Adams, R. J./Wilson, M./Wang, W. (1995). The multidimensional random coefficients multinomial logit model. In: Applied Psychological Measurement, 21, S. 1-23.

Adams, R. J./Wu, M. L. (2007). The mixed-coefficients logit model: A generalized form of the Rasch model. In: von Davier, M./Carstensen, C. H. (Hrsg,): Multivariate and mixture distribution Rasch models. New York, Berlin: Springer, S. 57-75.

Adams, R. J./Wu, M. L./Carstensen, C. H. (2007). Application of multivariate Rasch models in international large-scale educational assessments. In: von Davier, M./Carstensen, C. H. (Hrsg.): Multivariate and mixture distribution Rasch models. New York, Berlin: Springer, S. 271-280.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. In: Sociological Methodology, 20, S. 93-114.

Beaton, E./Allen, N. (1992). Interpreting scales through scale anchoring. In: Journal of Educational Statistics, 17, S. 191-204.

Beck, B./Klieme, E. (2007). Sprachliche Kompetenzen – Konzepte und Messung. Weinheim: Beltz.

Bollen, K. A. (1989). Structural equations with latent variables. New York: Wiley.

Brick, M./Morganstein, D./Valliant, R. (2000). Analysis of Complex Sample Data Using Replication. Rockville: Westat.

Campbell, D. T./Kenny, D. A. (2003). A primer on regression artifacts. New York, London: Guilford Press.

Cohen, J. ([2]1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. In: Psychological Bulletin, 112, S. 155-159.

Fay, R. E. (1989). Theory and application of replicate weighting for variance calculations. In: Proceedings of the Section on Survey Research of the American Statistical Association, S. 212-217.

Ganzeboom, H. B. G./de Graaf, P. M./Treiman, D. J./de Leeuw, J. (1992): A standard international socio-economic index of occupational status. In: Social Science Research 21, S. 1-56.

Goldhammer, F./Hartig, J. (2007). Interpretation von Testresultaten und Testeichung. In Moosbrugger, H./Kelava, A. (Hrsg.): Test- und Fragebogenkonstruktion. Berlin: Springer, S. 165-192.

Harsch, C/Schröder, K. (2007). Textrekonstruktion: C-Test. In: Beck, B./Klieme, E. (Hrsg.): Sprachliche Kompetenzen – Konzepte und Messung. Weinheim: Beltz, S. 212-225.

Hartig (2007). Skalierung und Kompetenzniveaus. In Beck, B./Klieme, E. (Hrsg.): Sprachliche Kompetenzen – Konzepte und Messung. Weinheim: Beltz, S. 83-99.

Hartig, J./Frey, A. (2006). Using plausible values from multidimensional irt models to estimate change in large scale assessments. Paper presented at the 15th international meeting of the Psychometric Society in Montréal, June 14–17 2006.

Hartig, J./Jude, N. (2005). Effects of different estimators for student proficiencies on multidimensional multilevel structures. Paper presented at the Fifth International Amsterdam conference on Multilevel Analysis, Amsterdam, March 21-23 2005.

Hartig, J./Jude, N. (2006). Einschätzung von Unterrichtswirksamkeit auf Basis von (nur) zwei Messzeitpunkten. Vortrag auf der 68. Tagung der Sektion Empirische Bildungsforschung der DGfE in München vom 10. bis 13. September 2006.

Hartig, J./Kühnbach, O. (2006). Schätzung von Veränderung mit Plausible Values in mehrdimensionalen Rasch-Modellen. In: Ittel, A./Merkens, H. (Hrsg.): Veränderungsmessung und Längsschnittstudien in der Erziehungswissenschaft. Wiesbaden: Verlag für Sozialwissenschaften, S. 27-44.

Heller, K. A./Perleth, C. (2000). Kognitiver Fähigkeitstest für 4. bis 12. Klassen. Revision. Göttingen: Beltz.

Helmke, A./Hosenfeld, I. (2004): Vergleichsarbeiten – Standards – Kompetenzstufen: Begriffliche Klärungen und Perspektiven. In: Jäger, R.S./Frey, A. (Hrsg.): Lernprozesse, Lernumgebung und Lerndiagnostik. Wissenschaftliche Beiträge zum Lernen im 21. Jahrhundert. Landau: Verlag Empirische Pädagogik.

Hox, J. (2003). Multilevel analysis: Techniques and applications. Mawah: Lawrence Erlbaum.

Jude, N./Hartig, J. (2006). Modelle zur Vorhersage von Leistungszuwächsen bei der Untersuchung von Unterrichtseffekten. Vortrag auf dem 45. Kongress der Deutschen Gesellschaft für Psychologie in Nürnberg vom 17. bis 21. September 2006.

Klieme, E./Baumert, J./Köller, O./Bos, W. (2000). Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In: Baumert, J./Bos, W./Lehmann, R. (Hrsg.): TIMSS/III. Dritte internationale Mathematik- und Naturwissenschaftsstudie. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit. Opladen: Leske + Buderich, S. 85-133.

Meiser, T./Stern, E./Langeheine, R. (1998). Latent Change in Discrete Data: Unidimensional, Multidimensional, and Mixture Distribution Rasch Models for the Analysis of Repeated Observations. In: Methods of Psychological Research Online, 3, S. 75-93.

Mislevy, R J./Beaton, A. E./Kaplan, B./Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of responses. In: Journal of Educational Measurement, 29, S. 133-161.

Muthén, L.K./ Muthén, B.O. (1998-2007). Mplus User's Guide. Fourth Edition. Los Angeles, CA: Muthén & Muthén

Nold, G../Rossa, H. (2007). Sprachbewusstheit. In: Beck, B./Klieme, H. (Hrsg.): Sprachliche Kompetenzen – Konzepte und Messung. Weinheim: Beltz, S. 226-244.

OECD (2001). Knowledge and Skills for Life. First Results from the OECD Programme for International Student Assessment (PISA) 2000. Paris: OECD.

OECD (2004). Learning for Tomorrow's World – First Results from PISA 2003. Paris:

Raghunathan T. E./Solenberger P. W./Hoewyk J. V. (2002). IVEware: Imputation and Variance Estimation Software. Installation Instructions and User Guide. Survey Research Center, Institute for Social Research, University of Michigan.

Rauch, D./Hartig, J. (2007). Interpretation von Testwerten in der IRT. In Moosbrugger, H./Kelava, A. (Hrsg.): Test- und Fragebogenkonstruktion. Berlin: Springer, S. 240-250.

Raudenbush, S./Bryk, A./Congdon, R. (2004). HLM 6 for Windows. Chicago, IL: Scientific Software.

Rost, J. (2004). Lehrbuch Testtheorie – Testkonstruktion. 2. überarbeitete und erweiterte Auflage. Bern, Göttingen: Huber.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. London: Chapman & Hall.

Schafer, J. L./Olsen, M. (1998). Multiple imputation for multivariate missing-data problems: a data analyst's perspective. In: Multivariate Behavioral Research 33, S. 545-571.

Schermelleh-Engel, K./ Moosbrugger, H. (2002). Beurteilung der Modellgüte von Strukturgleichun gsmodellen. Frankfurt: Arbeiten aus dem Institut für Psychologie der J. W. Goethe-Universität, Heft 4/2002.

Snijders, T.A.B./Bosker, R.J. (1994). Modeled variance in two-level models. In: Sociological Methods and Research, 22, S. 342-363.

Walter, O. (2006). Kompetenzmessung in den PISA-Studien. Simulationen zur Schätzung von Verteilungsparametern und Reliabilitäten. Lengerich: Pabst.

Westat (2000). WesVar 4.0 User's guide. Rockville: Westat.

Wilson, M. R. (2003). On choosing a model for measuring. Methods of Psychological Research Online, 8, 1-22.

Wolf, A. (2006). Shorter Tests Through the Adaptive Use of Planned Missing Data in Sampling Designs. Unpublished PhD Thesis, Friedrich-Schiller University Jena.

Wu, M. L./Adams, R. J./Wilson, M. R. (1998). ConQuest: Generalized item response modelling software. Melbourne: Australian Council for Educational Research.