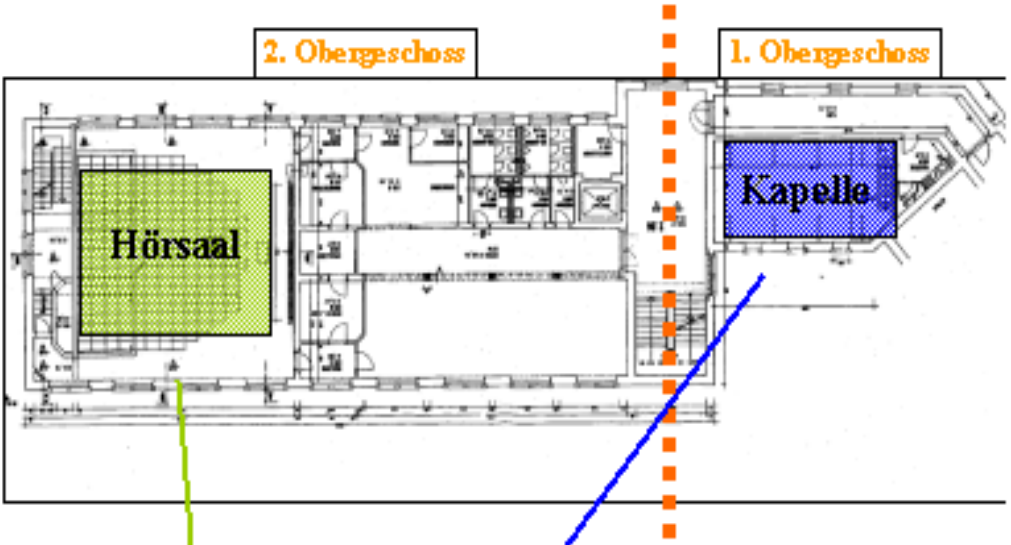
10. TAGUNG DER FGME IN BAMBERG

Fachgruppe für Methoden und Evaluation

21.-23.September 2011

Marcushaus M3, Markusplatz 3
Integrierte Ansicht des 1. und 2. Obergeschosses



1. Obergeschoss:
Kapelle
M3/126N
(K)

2. Obergeschoss:
Hörsaal
M3/232N
(H)

10. Tagung der FGME in Bamberg

Fachgruppe Methoden und Evaluation

21.–23. September 2011

GEFÖRDERT VON

DFG

DGPs Deutsche Gesellschaft für Psychologie

Fachgruppe für Methoden und Evaluation
Vorsitzender: Christof Schuster, Justus-Liebig-Universität Gießen

Impressum

Professur für Psychologie mit dem Schwerpunkt Methoden empirischer Bildungsforschung
Fakultät für Humanwissenschaften
Otto-Friedrich-Universität Bamberg
Wilhelmsplatz 3
96047 Bamberg

Organisationsteam: Professur für Psychologie mit dem Schwerpunkt Methoden empirischer Bildungsforschung
Claus H. Carstensen
Steffi Pohl
Eunike Wetzel
Claudia Tischler
Sophie-Catrin Schneider

Druck: Hausdruckerei der Universität Bamberg

Inhaltsverzeichnis

Preisverleihungen

Gustav A. Lienert-Preis

Gestiftet vom Springer-Verlag

Preisgeld: 1.500€

Diplomarbeitspreis

Gestiftet von der Fachgruppe für Methoden und Evaluation

Preisgeld: 500€

Die Preise werden am Mittwoch, den 21. September 2011, beim Gesellschaftsabend auf der Altenburg verliehen.
Der/die Preisträger/in des Gustav A. Lienert-Preises hält seinen/ihren Vortrag am Donnerstag, den 22. September 2011, um 14.00 Uhr im Hörsaal 232N (H).

Allgemeine Informationen

Tagungsräume

Die Tagungsräume befinden sich im Unigebäude Marcushaus (M3), Markusplatz 3, 96047 Bamberg.

Tagungsbüro

Das Tagungsbüro ist ab Mittwoch, 9.00 Uhr geöffnet. Es befindet sich im 2. Stock des Marcushauses im Foyer (M3/227N). Hier können Sie sich zur Tagung anmelden und erhalten Ihre Tagungsunterlagen und Namensschilder.
Bitte geben Sie Ihre Namensschilder am Ende der Tagung im Tagungsbüro ab. Vielen Dank.

Öffnungszeiten des Tagungsbüros

Mittwoch, 21. September 2011, 9.00-18.00 Uhr
Donnerstag, 22. September 2011, 8.30-18.00 Uhr
Freitag, 23. September 2011, 8.30-16.00 Uhr

Garderobe

Foyer (M3/ 227N), Marcushaus

Internetzugang

W-LAN im Marcushaus:
Die Zugangsdaten erhalten Sie zusammen mit Ihren Tagungsunterlagen.

Stationäre Rechner mit Internetzugang befinden sich im Marcushaus in der Bibliothek (Teilbibliothek 2) im Erdgeschoss sowie in den CIP-Pools im Kellergeschoss (M3/K10 und M3/K19N)
Die Zugangsdaten erhalten Sie zusammen mit Ihren Tagungsunterlagen.

Anreise zum Tagungsort

Buslinien:
Das Marcushaus erreichen Sie mit den Buslinien 906 (Haltestelle Markusplatz), 910 (Haltestelle Markusstraße) und 916 (Haltestelle Markusstraße).

Fußweg:
Das Marcushaus erreichen Sie vom Bahnhof in 25 Minuten und vom ZOB (Zentraler Omnibusbahnhof) in 10 Minuten.

Taxi-Zentralruf Bamberg:
0951/19410

Pausenversorgung

Während der im Tagungsprogramm angegebenen Kaffeepausen können Sie im Foyer Gebäck und Obst zu sich nehmen. Getränke sind immer erhältlich.

Die Mensa (Austraße 37) ist zum Mittagsessen von 11.30-14.00 Uhr geöffnet.

Empfehlenswert und nah am Tagungsort sind folgende Restaurants und Cafés:
Pizzeria Cuatro Gattos, Kapuzinerstraße 34
Gasthof Griesgarten, Untere Sandstraße 19
Spaghetteria Orlando, Jesuitenstraße 3
Café Esspress, Austraße 33
Inder Swarg, Frauenstraße 2

Ausflugtipps

Berühmt sind das Alte Rathaus in der Regnitz, der Bamberger Dom mit der angrenzenden Alten Hofhaltung und der Neuen Residenz sowie die Sandstraße.

Rahmenprogramm

Dienstag, 20. September 2011
Nachtwächterführung um 20.45
Treffpunkt: Tourist Information, Geyerswörthstraße 5

Mittwoch, 21. September 2011
Der diesjährige Gesellschaftsabend findet im historischen Rittersaal auf der Altenburg statt. Dort werden die Preise verliehen. Das Busshuttle zur Altenburg fährt um 19.00 Uhr an der Promenandenstraße (ZOB) ab, ein weiterer Zustieg ist an der Schranne um ca. 19.05 Uhr.

Donnerstag, 22. September 2011
Die Preisträger halten Ihre Vorträge um 14.00 Uhr.
Um 17.00 Uhr versammelt sich die Fachgruppe

Workshops Dienstag, 20. 9.

9-13 Uhr: Erich Neuwirth
RExcel: Anspruchsvolle Statistik kombiniert mit Tabellenkalkulation

9-13 Uhr: Steffi Pohl
Längsschnittmodellierung mit M*plus*

14-18 Uhr: Matthias von Davier & Claus H. Carstensen
IRT-Modelle mit dem HGDM

14-18 Uhr: Philipp Mayring
Qualitative und quantitative Textanalyse in den Sozialwissenschaften

Tagungsprogramm

Mittwoch, 21.09.2011

	Hörsaal M3/232N (H)
10:00	**Eröffnung**
10:30	**Modellierung von Veränderungen** Moderator: Axel Mayer
	Stemmler & Petersen: Die Anwendung von Latenten Wachstumskurvenmodelle zur Erforschung des Problemverhaltens bei jungen Frauen
	Fiege & Mayer: Die Schätzung von True-Change-Modellen als Multilevel Modelle
	Gasimova, Robitzsch, Wilhelm & Hülür: Hierarchisches Random-Effects-Modell für längsschnittliche Daten
11:30	**Keynote**
	Matthias von Davier: Some applications of multilevel diagnostic models to large scale assessment data
12:30	**Mittagspause**
14:00	**Symposium: Methodische Innovationen in der Testdiagnostik** Organisatoren: Johannes Hartig & Andreas Frey
	Goldhammer & Kröhne: Kontrolle individueller Bearbeitungsgeschwindigkeit zur Standardisierung von Leistungsmessungen
	Buchholz & Hartig: Latente Interaktionen in mehrdimensionalen IRT-Modellen
	Strobl: Risiken und Nebenwirkungen optimal-selektierter Statistiken am Beispiel der modellbasierten rekursiven Partitionierung von Rasch-Modellen
	Frey: Auswirkungen unterschiedlich balancierter Testheftdesigns auf Effizienz und Erwartungstreue am Beispiel von PISA 2006
15:20	**Kaffeepause**
15:50 – 17:10	**IRT-Modelle** Moderator: Christian Tarnai
	Walter: Penalisierte Spline Regression für Item Response Modelle
	Maier: Eine Alternative zu nominalen Antwortformaten und deren Analyse: Die Dirichletverteilung und Dirichletregressionsmodelle in R mit DirichletReg
	Heine, Tarnai & Hartmann: Eine Methode zur Parameterschätzung im Rasch-Modell bei fehlenden Werten
	Schuster & Berthold: GTT: Eine Software zur Anpassung logistischer Testmodelle
19:00 19:30	**Abfahrt zur Altenburg** **Gesellschaftsabend auf der Altenburg** **Preisverleihung**

Kapelle M3/126N (K)

Alternative Testmodelle
Moderator: Christof Schuster

Debelak: Evaluation eines Algorithmus zur modellgeleiteten Itemgruppierung

Borgstede: Psychologisches Testen als fuzzy-logische Inferenz

Plieninger, Meiser & Mienert: Modellierung von Item-Multidimensionalität und personenspezifischer Veränderung im Rahmen von Nonlinear Mixed Models

Modellierung fehlender Werte
Moderator: Claus H. Carstensen

Hohensinn & Kubinger: Wie wirkt sich die Verrechnung von ausgelassenen Items auf den Item Fit im Rasch Modell aus?

Thoemmes & Rose: Nicht alle Hilfsvariablen für fehlende Werte sind hilfreich

Wetzel & Carstensen: Äquivalenz der Testhefte von PISA 2000 und PISA 2009

Roppelt & Penk: Schätzfehler durch heterogene Varianzen bei Analysen mit Plausible Values

Anwendungsorientierte methodische Forschung I
Moderatorin: Karin Schermelleh-Engel

Colonius: Towards a measure of audiovisual integration efficiency

Zinn: Itemselektion in komplexen Zusammenhängen

Vetterlein & Sengewald: Mittelwerte verstehen! Rezeption von Ergebnisberichten im Hochschulsektor

Pfeiffer, Hagemann & Backenstrass: Eine neue Methode zur Schätzung der Varianzüberschneidung zwischen der Kurz- und der Langform eines psychologischen Tests

Donnerstag, 22. September 2011

	Hörsaal M3/232N (H)
09:00	**Symposium: Items, responses, and theory** Organisatoren: Steffi Pohl & Jan Marten Ihme
	Hartig & Buchholz: Analyse von Itempositionseffekten und inter-individuellen Unterschieden in diesen Effekten in einem Mehrebenen-IRT-Modell
	Ihme: Ein anderer Blick auf DIF *oder* Eine Strategie für Schiffe versenken
	Rose & von Davier: IRT Modelle für nonignorable Missing Data bei komplexer Dimensionalität
	Pohl, Gräfe & Hardt: Ignorability & Modellierung von fehlenden Werten in Kompetenztests
	Robitzsch: Berechnung von Standardfehlern in Large-Scale Assessments auf Grund von Item Sampling für Ländermittelwerte, originale und marginale Trendschätzungen
	Carstensen, Lankes & Steffensky: Homogenität von Kompetenzverläufe über drei Messzeitpunkte am Beispiel der Studie zu Naturwissenschaftlichen Kompetenzentwicklung im Elementarbereich – SnaKE
11:00	**Kaffeepause**
11:30	**Keynote**
	Cees Glas: IRT Modeling with non-ignorable Missing Data
12:30	**Mittagspause**
14:00	**Vorträge der Preisträger**
15:20	**Kaffeepause**
15:50	**IRT-Modellgeltungstests** Moderatorin: Carolin Strobl
	Alexandrowicz & Draxler: Tücken des Bootstrap
	Koller & Hatzinger: Nichtparametrische Tests für das Rasch Modell bei kleinem Stichprobenumfang: Eine empirische Poweranalyse
	Spoden, Fleischer, Zischka & Leutner: Hypothesentests bei Rasch Personen-Fit-Statistiken - Eine Alternative zum konventionellen Monte-Carlo-Verfahren
17:00	**Versammlung der Fachgruppe**
19:00	**Abendessen im Luitpold**

Kapelle M3/126N (K)
Modellierung latenter Variablen Moderator: Markus Bühner
Könen, Preckel & Brunner: Alternative Verwendung eines CS-C(M−1) Modells: Erfassung von Domänenspezifität durch Residualfaktoren
Mutz: Die räumliche Dimension in der Psychometrie: Ein latentes 'Geopress-State'-Modell
Heene & Maraun: Faktorstrukturinvarianz = Faktorinvarianz: Ein weit verbreiteter Irrglaube
Nestler: Eine Monte-Carlo Studie zum Vergleich der Genauigkeit des ULS-, DWLS- und PIV-Schätzers bei dichotomen konfirmatorischen Faktorenanalysen
Sengewald & Vetterlein: Einfluss der hierarchischen Struktur auf den Modellfit für Instrumente der Lehrevaluation
Böhme H.: Anwendung von Latente-Variablen-Modellen zur Testgütekriterienbestimmung bei Testverfahren zur Erfassung nominaler Konstrukte
Anwendungsorientierte methodische Forschung II Moderator: Thorsten Meiser
Baghaei: Validation of a multidimensional willingness to communicate scale
Poinstingl, Herzberg, Brähler & Kubinger: Psychometrische Untersuchung der Hopkins Symptom Checklist 25 (HSCL-25) mit dem Partial Credit Modell
Böhme K. & Schipolowski: Test- und Raterdesigns für das Assessment und die Kodierung freier Schreibaufgaben

Freitag, 23.September 2011

	Hörsaal M3/232N (H)
09.00	**Symposium: Kausale Effekte: Fortschritte in Theorie und Anwendung** Organisator: Benjamin Nagengast
	Mayer, Kirchmann & Steyer: Behandlungseffekte für Behandelte (Effects on the treated) in kausalen Mediationsmodellen
	Thoemmes, Mayer & Steyer: Vergleich von Kausalitätsbedingungen
	Hahn: Vergleich von Permutations- und parametrischen Tests bei Treatment-Kovariaten-Interaktion
	Nagengast, Marsh, Chiorri & Hau: Adjustierungsverfahren für geordnet kategoriale Treatments: Ein theoretischer Vergleich und eine empirische Anwendung
	Kröhne, Hartig & Klieme: Beurteilung der Wirksamkeit bilingualen Unterrichts: Vergleich von Strategien zur Identifikation und von Verfahren zur Schätzung kausaler Effekte in Large-Scale Assessments am Beispiel einer Re-Analyse einer Teilstichprobe der DESI-Studie
	Steyer: Diskussion
11:00	**Kaffeepause**
11:30	**Keynote**
	Tenko Raykov: On Missing Data Mechanisms and Analysis of Incomplete Data Sets in Behavioral, Social, and Biomedical Research
12:30	**Mittagspause**
14:00	**Nichtlineare Strukturgleichungsmodellierung** Moderator: Andreas Klein
	Werner & Schermelleh-Engel: Analyse von Moderatoreffekten mit Strukturgleichungsmodellen oder Partial Least Squares: Konsequenzen des Orthogonalisierens von Produkttermen
	Gerhard, Schermelleh-Engel, Werner & Moosbrugger: Chi-Quadrat Differenztest in nichtlinearen Strukturgleichungsmodellen
	Schermelleh-Engel, Werner, Gäde, Gerhard & Moosbrugger: Zum Problem der korrekten Standardisierung von nichtlinearen Strukturgleichungsmodellen
	Kelava & Nagengast: Ein neues Verfahren zur Schätzung von latenten nichtlinearen Effekten bei nicht-normalverteilten latenten Prädiktoren
15:20	**Kaffeepause**
15:50	**Mediationsanalysen** Moderator: Helfried Moosbrugger
	Klein & Brandt: Conceptual difficulties of conventional ways of modeling mediation
	Brandt & Klein: Identification of mediator variables in randomized longitudinal trials
	Rentzsch: Die Kombination von Social Relations- und Multilevel Mediation Analysen im Klassenkontext
16:50	**Tagungsende**

Kapelle M3/126N (K)

Symposium: Psychometrische Probleme bei Antwort- und Itemformaten
Organisator: Moritz Heene

Heene, Zech, Hilbert & Bühner: Itempolung und seine Effekte auf die Qualität von Ratingskalen: Eine Überprüfung mithilfe des Generalized Partial Credit Modells

Zech, Heene, Hilbert, Bollmann & Bühner: Auswirkungen von vague quantifiers auf die faktorielle Struktur von Fragebogen

Hilbert, Heene, Zech, Ziegler & Bühner: Dichotom, ordinal oder analog? Wie unterschiedliche Antwortformate psychometrische Eigenschaften von Fragebogenitems bestimmen

Ziegler, Danay, Gammon & Griffith: Response Set oder Faking-Stil? Eine Analyse von Bewerberdaten

Einzelbeitrag

Yousfi: A framework for test equating based on the concepts of equity and causal effects

Antwortstile
Moderator: Rainer Alexandrowicz

Keller & Alexandrowicz: Untersuchungen zu Rasch-Homogenität und response sets im Beck Depressions-Inventar Revision (BDI-II)

Heine, Langmeyer, Tarnai & Hartmann: Kontrolle von Antwortstilen durch die Anwendung von Raschmodellen und ihre Auswirkung auf Skaleninterkorrelationen

Wetzel & Carstensen: Zusammenlegung von Antwortkategorien bei der Untersuchung von Antwortstilen

Meiser & Böckenholt: IRT-Analyse von Traitausprägung und Antwortstilen in Ratingdaten

Klassifikationen
Moderator: Alexander Robitzsch

George, Robitzsch & Groß: Analyse individueller Klassifikationen in verallgemeinerten DINA-Modellen

Bräker, Huber & Soellner: Clusteranalyse oder Latent-Class-Analyse? Ein Vergleich der Verfahren am Beispiel von Alkoholkonsummustern Jugendlicher in Europa

Tagungsbeiträge

Keynote-Speaker

Mi., 21.09., Keynote: Matthias von Davier, 11.30-12.30 Uhr, Hörsaal M3/232N (H)

Some applications of multilevel diagnostic models to large scale assessment data

Matthias von Davier

Educational Testing Service
Princeton, NJ, USA

Data collected in national and international educational surveys are hierarchically organized. Students are sampled from within schools, and schools are sampled within states or countries. Several methods of accounting for this multilevel structure have been developed for a variety of psychometric models. Examples are multilevel IRT (e.g. Fox and Glas, 2003) or hierarchical extensions of latent class analysis (Vermunt, 2003). For diagnostic classification approaches, the general diagnostic model (von Davier & Yamamoto, 2004) has been extended to a hierarchical general diagnostic model (HGDM, von Davier, 2007).

In this talk, the HGDM and some recent applications of this approach will be presented. The studies range from linking cognitive scales in international assessments (Oliveri & von Davier, 2011) to measuring growth in a national extension of an international study (von Davier, Xu & Carstensen, 2010) to studies on scaling (von Davier, Naemi & Roberts, 2011) and the imputation of missing background data (Xu & von Davier, 2011). Some direction for future extensions will be discussed.

Do., 22.09., Keynote: Cees Glas, 11.30-12.30 Uhr, Hörsaal M3/232N (H)

IRT modeling with non-ignorable missing data

Cees Glas

Department of Research Methodology, Measurement, and Data Analysis
Faculty of Behavioral Sciences
University of Twente
Enschede, the Netherlands

Missing data can cause bias in parameter estimates and lead to misleading conclusions. This is especially true when the so-called “ignorability principle” defined by Donald Rubin is violated. Roughly speaking, this principle entails that all variables related to the causes of the missingness should be observed. This presentation focuses on item-non-response. If skipping items or not-reaching of items is related to the construct which is the target of the measurement, it is plausible that the “ignorability principle” may be violated. An example is skipping of difficult items by low-ability children.

In this presentation, it is shown how the bias in the estimates of item and person parameters of an IRT model can be adjusted by adding a secondary IRT model for the missing data indicator. Three examples of such a secondary model are given. The first example shows how skipping of difficult items by low-ability children can be modeled using the Rasch model. In the second example, a sequential IRT model is used to model missing responses at the end of a speeded test. In the third example it is shown how a single-peaked IRT model can be used to adjust for differential choice of examination topics in secondary education. All three examples are applied to both simulated and real data. An often neglected topic in IRT modeling is the evaluation of model fit. Therefore, the second part of this presentation will be devoted to methods to evaluate the fit of the proposed models. Two decades ago, the presenter has proposed the Lagrange multiplier test as a general framework for the evaluation of specific assumptions of IRT models, such as the assumption regarding the shape of the response function, local independence, parameter invariance and distributional assumptions regarding the person parameters. It will be shown how this methodology can be adopted to testing the fit of the IRT models for non-ignorable missing data presented here.

Fr., 23.09., Keynote: Tenko Raykov, 11.30-12.30 Uhr, Hörsaal M3/232N (H)

On Missing Data Mechanisms and Analysis of Incomplete Data Sets in Behavioral, Social, and Biomedical Research

Tenko Raykov

Measurement & Quantitative Methods
Michigan State University
East Lansing, MI, USA

Missing data pervade empirical research in psychology and the behavioral, social, and biomedical disciplines. This talk shows that contrary to statements and implications from a considerable body of widely circulated and cited literature, the missing data mechanism routinely referred to as missing completely at random (MCAR) is not testable via examination for distributional differences between groups with observed and with missing data. A discussion is provided, from a formal logic standpoint, of the distinction between necessary conditions and sufficient conditions. It is shown that lack of group distributional differences is not sufficient for MCAR, and it is argued that the importance of MCAR has been frequently overrated in empirical behavioral, social, and biomedical research. A multiple testing approach to examining an incomplete data set for not being MCAR is outlined. A latent variable method discussed in Raykov (2001, *Structural Equation Modeling*) is recommended when fitting models containing covariates with missing values using maximum likelihood in the presence of auxiliary variables.

TAGUNGSBEITRÄGE

Mittwoch

Mi., 21.09., Modellierung von Veränderungen, 10.30-10.50 Uhr, Hörsaal M3/232N (H)

Die Anwendung von Latenten Wachstumskurvenmodelle zur Erforschung des Problemverhaltens bei jungen Frauen

Mark Stemmler[1]
Anne C. Petersen[2]

[1]Institut für Psychologie
Friedrich-Alexander-Universität Erlangen-Nürnberg
Bismarckstraße 1
91054 Erlangen

mark.stemmler@psy.phil.uni-erlangen.de

[2]Center for Human Growth and Development (CHGD), University of Michigan, 300 North Ingalls, 10th Floor, Ann Arbor, MI 48109-5406, United States

Anhand der Daten der Adolescent Mental Health Study (AMHS; Petersen, 1984), die einen Zeitraum von 10 Jahren umfassen, soll die Beziehung zwischen der psychischen Anpassung in der Adoleszenz, die insgesamt fünf Mal erhoben wurde und dem Problemverhalten im jungen Erwachsenenalter untersucht werden. Das Problemverhalten umfasst Drogengebrauch (legale und illegale Drogen) sowie Normverletzendes Verhalten (z.B. Vandalismus, Diebstahl). Die längsschnittliche Stichprobe bestand aus n = 133 Mädchen. Latente Wachstumskurvenmodelle (LGCM; McArdle, 2009) auf der Basis von Strukturgleichungsmodellen werden mit einander verglichen. LGCM wird verwendet um intra- und interindividuelle Veränderung zu erfassen. Es zeigte sich, dass die psychische Anpassung der Mädchen, gemessen als Selbstkonzept bezüglich Copingverhalten, Beliebtheit bei Gleichaltrigen und guten Schulleistungen mit dem Problemverhalten im jungen Erwachsenenalter in Beziehung steht. So hatte Drogenkonsum einen signifikanten negativen Einfluss auf den Slope von psychischer Anpassung. Normverletztendes Verhalten zeigte eine positive Beziehung zum Drogenkonsum. Verschiedene latente Wachstumskurven werden mit einander verglichen und danach beurteilt inwieweit sie in der Lage sind unbeobachtete Heterogenitäten zu erklären.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. Annual Review of Psychology, 60, 577-605.

Petersen, A. C. (1984). The Early Adolescence Study: An overview. Journal of Early Adolescence, 4, 103-106.

Die Schätzung von True-Change-Modellen als Multilevel Modelle

Christiane Fiege
Axel Mayer

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl für Methodenlehre und Evaluationsforschung
Am Steiger 3, Haus 1
07743 Jena

christiane.fiege@uni-jena.de

axel.mayer@uni-jena.de

Multilevel Modelle (MLM) ermöglichen die Analyse von Forschungsdesigns mit einer Mehrebenenstruktur, in denen die Untersuchungseinheiten innerhalb größerer Einheiten (bspw. Schüler in Schulen) genestet sind. Sowohl Bauer (2003) als auch Curran (2003) widmeten sich bereits der Frage, ob und wie Multilevel Modelle (MLMs) in Strukturgleichungsmodelle (SEMs) überführbar sind. Beide Autoren stellen die konzeptuelle Äquivalenz beider Ansätze sowie die Schätzung von MLMs als SEMs dar.
Im Rahmen dieses Beitrags betrachten wir die entgegengesetzte Richtung, d.h. die Darstellung von SEMs als MLMs. Wir präsentieren einen Vergleich der beiden Ansätze im Kontext von True-Change-Modellen (TCMs) zur Modellierung latenter Veränderungsvariablen (Steyer, Partchev & Shanahan, 2000). Zwar haben hier SEMs klare Vorteile gegenüber MLMs in Bezug auf globale Modellgeltungskontrollen sowie der Testung spezifischer Implikationen der TCMs mittels Gleichheitsrestriktionen (bspw. die Invarianz des Messmodells). Umgekehrt liegt der Vorteil der Reparametrisierung dieser Modelle als MLMs darin, dass auf diese Weise TCMs leichter auf Daten mit einer Mehrebenenstruktur angewendet werden können. Zudem können Prädiktoren der latenten Veränderungsvariablen auf verschiedenen Ebenen modelliert werden.
In diesem Beitrag wird die Schätzung von True-Change-Modellen als MLMs dargestellt. Dazu werden zwei Versionen der TCMs – das Baseline-Modell und das Neighbor-Modell – betrachtet. Schließlich wird die Implementierung beider Ansätze anhand eines Datenbeispiels aus der Bildungsforschung illustriert. Möglichkeiten und Grenzen beider Ansätze werden diskutiert.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. Journal of Educational and Behavioral Statistics, 28, 135-167.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? Multivariate Behavioral Research, 38, 529-569.

Steyer, R., Partchev, I. & Shanahan, M. J. (2000). Modeling true intraindividual change in structural equation models: The case of poverty and children's psychosocial adjustment. In T. D. Little, K. U. Schnabel & J. Baumert (Hrsg.), Modeling longitudinal and multilevel data (S. 109-126). Mahwah, NJ: Lawrence Erlbaum Associates.

Mi., 21.09., Modellierung von Veränderungen, 11.10 -11.30 Uhr, Hörsaal M3/232N (H)

Hierarchisches Random-Effects-Modell für längsschnittliche Daten

Fidan Gasimova[1]
Alexander Robitzsch[2]
Oliver Wilhelm[1]
Gizem Hülür[1]

[1]Universität Ulm, Institut für Psychologie und Pädagogik
Albert-Einstein-Allee 47
89069 Ulm

fidan.gasimova@uni-ulm.de

[2]Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens Alpenstraße 121
5020 Salzburg

a.robitzsch@bifie.at

Dieser Beitrag befasst sich mit Modellierungstechniken für intensiv-längsschnittliche Daten. Im gegebenen Fall sind dies Schülerleistungen in Deutsch und Mathematik sowie Arbeitsgedächtnisleistungen aus der „LUISE“ Studie (Längsschnittliche Untersuchung individueller schulischer Entwicklungsprozesse).für die wir adäquate Analysemethoden für viele, nicht äquidistante Messzeitpunkte und viele Individuen einsetzen. Im Zentrum des Vorhabens steht die simultane Betrachtung intra- und interindividueller Veränderungen in Deutsch, Mathematik und Arbeitsgedächtniskapazität.

Oravecz et al. (2009) sowie Oravecz und Tuerlinckx (2011) schlagen personenspezifische stochastische Differentialgleichungen erster Ordnung für die Analyse der längsschnittlichen Daten, das im Gegensatz zu einem latenten linearen Wachstumskurvenmodell interindividuelle Veränderungen, nicht nur in einem personenspezifischen Mittelwert und einem personenspezifischen Slope, sondern auch um personenspezifische Varianzen und Autokorrelationen erweitern. Die Personenparameter werden dabei als hierarchische zufällige Effekte in einem so genannten Random Effects Model spezifiziert.

Im vorgeschlagenen Modell wird zunächst angenommen, dass die Veränderungen über die Zeit linear und die Abstände zwischen den einzelnen Messungen für alle Personen äquidistant sind. Mit Hilfe einer Simulationsstudie wurde das Random Effects Model auf Parameter Recovery untersucht und die statistischen Eigenschaften Bias, RMSE und Coverage evaluiert. Die zur Simulation der Datensätze verwendeten Populationsparameter wurden aus der LUISE-Studie in einer Stichprobe von N=138 Personen und T=20 Messzeitpunkten gewonnen.

Oravecz, Z., Tuerlinckx, F., & Vandekerckhove, J. (2009). A hierarchical Ornstein-Uhlenbeck model for continuous repeated measurement data. Psychometrika, 74, 395-418.

Oravecz, Z., & Tuerlinckx, F. (2011). The linear mixed model and the hierarchical Ornstein-Uhlenbeck model: Some Equivalences and differences. British Journal of Mathematical and Statistical Psychology, 64, 134-160.

Mi., 21.09., Alternative Testmodelle, 10.30-10.50 Uhr, Kapelle M3/126N (K)

Evaluation eines Algorithmus zur modellgeleiteten Itemgruppierung

Rudolf Debelak

Schuhfried GmbH
Hyrtlstraße 45, 2340 Mödling, Österreich

debelak@schuhfried.at

Ein neuer Ansatz zur Identifikation von Itemclustern, für welche das Rasch-Modell (Rasch, 1960) gilt, wird beschrieben, anhand einer Simulationsstudie evaluiert und mit einer praktischen Anwendung veranschaulicht. Die im Vortrag beschriebene Methode basiert auf dem Prinzip einer hierarchischen Clusteranalyse. Zunächst wird der vorgestellte Ansatz mit verwandten Analysemethoden der modernen Testtheorie (z.B. Bartolucci, 2007) verglichen. Verschiedene Anwendungsmöglichkeiten der Methode werden diskutiert und anhand der Ergebnisse einer Simulationsstudie veranschaulicht. In dieser Simulationsstudie wird gezeigt, inwiefern die Methode geeignet ist, die Dimensionalität eines Itemsets abzuschätzen, indem die Anzahl von Clustern bestimmt wird, welche in dem untersuchten Itemset gefunden werden können. Diese Studie belegt dabei, dass die beschriebene Methode umso bessere Ergebnisse erzielt, je kleiner das untersuchte Itemset ist, je größer die Stichprobe ist, welche das Itemset bearbeitete, und je weniger die den Items zugrunde liegenden Fähigkeitsdimensionen miteinander korrelieren. Die Ergebnisse der beschriebenen Methode werden dabei mit den Ergebnissen einer Hauptkomponentenanalyse tetrachorischer Korrelationen verglichen. Es zeigt sich, dass keine der beiden Methoden der anderen generell überlegen ist.

In einer zweiten Studie wird die Anwendung der Methode an einem Datensatz von 298 Personen, welche die Intelligenztestbatterie IBF (Blum, Didi, Fay, Maichle, Trost, Wahlen, & Gittler, 2005) bearbeiteten, dargestellt. Es zeigt sich dabei, dass die von der Methode konstruierten Itemcluster verschiedenen klassischen Modelltests für das Rasch-Modell (Andersen, 1973) genügen.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38, 123-140.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. Psychometrika, 72, 141-157.

Blum, F., Didi, H.-J. , Fay, E., Maichle, U., Trost, G., Wahlen, H.-J., & Gittler, G. (2005). Intelligenz-Basis-Funktionen (IBF). Mödling: Schuhfried.

Mi., 21.09., Alternative Testmodelle, 10.50-11.10 Uhr, Kapelle M3/126N (K)

Psychologisches Testen als fuzzy-logische Inferenz

Matthias Borgstede

Technische Universität Braunschweig
Institut für Psychologie
Abteilung Psychologische Methodenlehre und Biopsychologie
Spielmannstraße 12a
38106 Braunschweig

m.borgstede@tu-braunschweig.de

Es wird ein neuartiger formaler Ansatz zur Beschreibung und Auswertung psychologischer Tests vorgestellt, welcher auf dem Prinzip der fuzzy-logischen Inferenz basiert. Im Gegensatz zur traditionellen psychometrischen Sichtweise werden psychologische Tests nicht als Messinstrumente verstanden, sondern als eine Möglichkeit des logischen Erschließens von Personenattributen. Tests werden dabei nicht als Indikatoren latenter Variablen betrachtet, sondern als logische Propositionen, welche semantisch mit bestimmten (möglicherweise vagen) Konstruktbegriffen zusammenhängen. Durch die Anwendung fuzzy-logischer Schlussregeln ergeben sich aus den Antwortmustern eines Tests entsprechende Konstruktausprägungen. Der so ermittelte Testscore wird jedoch nicht als Messwert verstanden, sondern als das Ausmaß, in welchem ein Konstruktbegriff auf eine Person zutrifft. Da dieser formal-logische Schluss unabhängig von der Frage nach der Verursachung der Testantworten ist, kann der vorgestellte Ansatz als komplementär zu IRT-Modellen angesehen werden. Er bietet somit eine Möglichkeit, psychologische Tests zu konstruieren und zu optimieren, selbst wenn wenig oder gar nichts über die latente Struktur eines Konstrukts bekannt ist.

Borsboom, D. (2005). Measuring the mind. Conceptual issues in contemporary psychometrics. Cambridge: Cambridge University Press.

Gulliksen, H. (1950). Theory of mental tests. New York: J. Wiley.

Kline, P. (1998). The new psychometrics. Science, psychology and measurement. London: Routledge.

Klir, G., & Yuan, B. (1995). Fuzzy sets and fuzzy logic. New Jersey: Prentice Hall.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental scores. Reading, MA: Addison–Wesley.

Rost, J. (2004). Lehrbuch Testtheorie, Testkonstruktion. Zweite, vollständig überarbeitete und erweiterte Auflage. Bern, Stuttgart, Wien: Verlag Hans Huber.

Zadeh, L. (1965). Fuzzy sets. Information and control, 8 (3), 338–353.

Mi., 21.09., Alternative Testmodelle, 11.10-11.30 Uhr, Kapelle M3/126N (K)

Modellierung von Item-Multidimensionalität und personenspezifischer Veränderung im Rahmen von Nonlinear Mixed Models

Hansjörg Plieninger
Thorsten Meiser
Malte Mienert

Lehrstuhl Psychologie II
Universität Mannheim
Schloss Ehrenhof
68131 Mannheim
hplienin@mail.uni-mannheim.de

Der vorliegende Beitrag basiert auf der Verknüpfung von Item-Response-Theorie und Multilevel-Modellen zu sog. nonlinear mixed models (De Boeck & Wilson, 2004) bzw. hierarchical generalized linear models (Kamata, 2001). Dabei werden die Items auf Level 1 als fixed effects betrachtet, während der Personenparameter auf Level 2 als random effect mit einer bestimmten Verteilungsannahme angesehen wird. Diese Modellierung ist zunächst äquivalent zum Rasch-Modell, bietet jedoch die Möglichkeit zu zahlreichen Erweiterungen. Beispielsweise können mittels random slopes personenspezifische Effekte von Itemkomponenten modelliert werden, Prädiktoren für die Item- oder Personenparameter können einfließen, oder es besteht die Möglichkeit für eine dritte Ebene im Fall von cluster samples. Außerdem stellt die Spezifikation als nonlinear mixed model einen gemeinsamen Rahmen für eine Vielzahl von IRT-Modellen dar.

Der beschriebene Ansatz wird auf die Zerlegung polytomer Items in einzelne dichotome Items (steps model; Verhelst, Glas, & de Vries, 1997) angewendet. Damit ist es möglich Multidimensionalität innerhalb der Items zuzulassen, indem für verschiedene thresholds unterschiedliche latente Variablen spezifiziert werden.

Die Kombination des nonlinear mixed model-Ansatzes mit dem steps model wird anhand längsschnittlicher Daten von über 600 Jugendlichen illustriert. Diese haben im Abstand von ca. einem Jahr jeweils 18 trichotome Items zu Entwicklungsaufgaben des Jugendalters bearbeitet.

Die Analyse bestätigt die Annahme, dass zwei verschiedene, aber korrelierte Traits auf die beiden Schwellenübergänge wirken. In einer Sequenz von Modellen wird illustriert und diskutiert, welche Vorteile der vorgestellte Ansatz bietet um beispielsweise Variabilität im Ausmaß der Veränderung - modelliert als random slope - oder den Einfluss von Prädiktoren zu untersuchen.

De Boeck, P., & Wilson, M. (Eds.). (2004). Explanatory item response models: A generalized linear and nonlinear approach. New York: Springer.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. Journal of Educational Measurement, 38, 79-93. doi:10.1111/j.1745-3984.2001.tb01117.x

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden, & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 123-138). New York: Springer.

Mi., 21.09., Symposium: Methodische Innovationen in der Testdiagnostik, 14.00-15.20 Uhr, Hörsaal M3/232N (H)

Symposium: Methodische Innovationen in der Testdiagnostik

Johannes Hartig[1]
Andreas Frey[2]

[1]Deutsches Institut für internationale Pädagogische Forschung (DIPF)

[2]Friedrich-Schiller-Universität Jena

Zielsetzung

Die Fortentwicklung der Methodologie zur Auswertung der mit standardisierten Testverfahren erhobenen Antwortdaten, wird international wie national maßgeblich durch die Anforderungen von Studien in der empirischen Bildungsforschung, insbesondere in groß angelegten Vergleichsstudien (Large-Scale-Assessments), beeinflusst. Ziel des Symposiums ist es, einen Einblick in aktuelle Methodenentwicklungen in verschiedenen Bereichen der Testdiagnostik im Kontext der empirischen Bildungsforschung zu geben. Der Schwerpunkt liegt dabei auf Modellen der Item-Response-Theorie (IRT). Die einzelnen Beiträge beschäftigen sich mit der Kontrolle von Bearbeitungsgeschwindigkeiten im Kontext computerbasierten Testens (Goldhammer & Kröhne), der Nutzung latenter Interaktionsterme zur Modellierung nicht-kompensatorischer Funktionen in mehrdimensionalen IRT-Modellen (Buchholz & Hartig), der Identifikation von Personengruppen mit unterschiedlichen Itemparametern mithilfe rekursiver Partitionierung (Strobl) sowie der Nutzung balancierter unvollständiger Testheftdesigns zur Kontrolle von Itempositionseffekten bei PISA (Frey).

Beiträge

1. Goldhammer, Frank & Kröhne, Ulf: Kontrolle individueller Bearbeitungsgeschwindigkeit zur Standardisierung von Leistungsmessungen
2. Buchholz, Janine & Hartig, Johannes: Latente Interaktionen in mehrdimensionalen IRT-Modellen
3. Strobl, Carolin: Risiken und Nebenwirkungen optimal-selektierter Statistiken am Beispiel der modellbasierten rekursiven Partitionierung von Rasch-Modellen
4. Frey, Andreas: Auswirkungen unterschiedlich balancierter Testheftdesigns auf Effizienz und Erwartungstreue am Beispiel von PISA 2006

Mi., 21.09., Symposium: Methodische Innovationen in der Testdiagnostik, 14.00-14.20 Uhr, Hörsaal M3/232N (H)

Kontrolle individueller Bearbeitungsgeschwindigkeit zur Standardisierung von Leistungsmessungen

Frank Goldhammer
Ulf Kröhne

Deutsches Institut für Internationale Pädagogische Forschung (DIPF)
Schloßstr. 29
60486 Frankfurt/Main

goldhammer@dipf.de

Aus testdiagnostischer Sicht stellt der Speed-Ability Tradeoff (SAT) ein grundlegendes Problem dar, insofern durch ihn die Vergleichbarkeit von leistungsbezogenen Messungen (z.B. von Fähigkeiten, Kompetenzen) beeinträchtigt werden kann. Individuelle Unterschiede in der gewählten Bearbeitungsgeschwindigkeit können in Leistungstests z.B. zu unterschiedlichen Fähigkeitsschätzungen führen, auch wenn sich die Personen in ihrer Fähigkeit bei gleicher Bearbeitungsgeschwindigkeit nicht unterscheiden würden (vgl. van der Linden, 2009). Mit vorliegender Studie soll durch Entwicklung und Einsatz einer speziellen computerbasiertem Testprozedur der SAT untersucht und auf individueller Ebene kontrolliert werden. Die Studie wendet dazu das experimentelle Response-Signal-Paradigma (Reed, 1973) an, um die individuelle Bearbeitungsgeschwindigkeit durch Vorgabe eines akustischen Reaktionssignals und eines begrenzten Reaktionszeitfensters zu kontrollieren. Die Aufgaben werden in einem Inner-Subjekt-Design ohne sowie mit unterschiedlich starken Zeitbegrenzungen administriert. Es wird gemäß SAT erwartet, dass sich mit Verkürzung der Zeit bis zum Reaktionssignal die Leistung verringert sowie die Leistungsvariabilität abnimmt. Es werden höhere Korrelationen zwischen den Bedingungen mit Zeitbegrenzung erwartet als zwischen der Bedingung ohne Zeitbegrenzung und einer Bedingung mit Zeitbegrenzung. Die Testprozedur wurde exemplarisch am Beispiel von visuellen Worterkennungsaufgaben sowie figuralen Diskriminationsaufgaben zur Erfassung exekutiver Aufmerksamkeit an einer Schülerstrichprobe erprobt. Die Manipulationskontrolle zeigt, dass individuelle Unterschiede in der Bearbeitungsgeschwindigkeit durch das Response-Signal-Paradigma verringert werden. Durch Testung eines mehrdimensionalen IRT-Modells mit Bedingungen als Dimensionen und invarianten Messmodellen kann gezeigt werden, dass gemäß SAT das gezeigte Fähigkeitsniveau für Worterkennung mit Zunahme der Geschwindigkeit abnimmt, und dass die Fähigkeitsvarianz zunächst zunimmt, um dann über die beiden extremsten Geschwindigkeitsbedingungen stark abzunehmen. Deutliche höhere Korrelationen zwischen den Bedingungen mit Zeitbegrenzung zeigen an, dass der SAT durch das Response Signal-Paradigma auf individueller Ebene kontrolliert werden kann. Die Ergebnisse sind generalisierbar auf Diskriminationsaufgaben zur Erfassung exekutiver Aufmerksamkeit.

Reed, A. V., 1973. Speed-accuracy tradeoff in recognition memory. Science 181, 574-576.
van der Linden, W.J. (2009). Conceptual issues in response-time modeling. Journal of Educational Measurement, 46, 247-272.

Mi., 21.09., Symposium: Methodische Innovationen in der Testdiagnostik, 14.20-14.40 Uhr, Hörsaal M3/232N (H)

Latente Interaktionen in mehrdimensionalen IRT Modellen

Janine Buchholz
Johannes Hartig

Deutsches Institut für Internationale Pädagogische Forschung (DIPF)
Schloßstraße 29
60486 Frankfurt am Main

buchholz@dipf.de

Während latente Interaktionen in der Klasse der Strukturgleichungsmodelle verbreitet sind, gibt es bisher nur sehr wenig Erfahrung damit innerhalb der mehrdimensionalen Item Response Theorie (MIRT). Dabei ist ihr Einsatz im Kontext der Leistungsdiagnostik theoretisch vielversprechend, zum Beispiel dann, wenn hohe Lösungswahrscheinlichkeiten erst durch das gemeinsame Vorliegen mehrerer Fähigkeiten zustande kommen sollen. Für solche non-kompensatorischen Beziehungen zwischen latenten Variablen wurden bereits MIRT-Modelle formuliert, in denen die Lösungswahrscheinlichkeiten als Produkt mehrerer logits modelliert werden. Simulationsstudien zeigen für diese Modelle jedoch sehr hohe Anforderungen an die Daten, neben großen Stichproben sind insbesondere „reine“ Items (d. h. Items, die nur von einer der Dimensionen beeinflusst werden) für jede der latenten Fähigkeiten notwendig. In einer Simulationsstudie wurde untersucht, unter welchen Bedingungen ein MIRT-Modell mit latenter Interaktion zuverlässige Schätzungen liefern kann. Dafür wurden dichotome Itemantworten unter Gültigkeit eines zweidimensionalen 2PL- Modells mit latenter Interaktion für einen Teil der Items simuliert. Es wurde untersucht, wie gut sich die Interaktionseffekte schätzen lassen, wenn für beide oder nur für eine der latenten Dimensionen reine Items als Indikatoren vorliegen. Variiert wurden weiterhin die Stichprobengröße und die Anzahl der Items pro Dimension. Sowohl Simulation als auch Modellschätzung erfolgten in Mplus. Die Ergebnisse zeigen, dass latente Interaktionseffekte auch dann geschätzt werden können, wenn auch mit geringerer Power, wenn für nur eine der beteiligten Dimensionen reine Items vorliegen.

Mi., 21.09., Symposium: Methodische Innovationen in der Testdiagnostik, 14.40-15.00 Uhr, Hörsaal M3/232N (H)

Risiken und Nebenwirkungen optimal-selektierter Statistiken am Beispiel der modellbasierten rekursiven Partitionierung von Rasch-Modellen

Carolin Strobl

Institut für Statistik
Ludwig-Maximilians-Universität München
Ludwigstr.33
80539 München

carolin.strobl@stat.uni-muenchen.de

Die modellbasierte rekursive Partitionierung kann zum Aufdecken von Parameter-Invarianzen in psychometrischen Modellen verwendet werden, wie z.B. zur Diagnose von Differential Item Functioning im Rasch-Modell. Dabei werden Personen-Kovariablen (z.B. Alter und Geschlecht) eingesetzt, um Gruppen von Personen mit unterschiedlichen Aufgaben-Parametern zu identifizieren. Ein Vorteil dieses Verfahrens gegenüber herkömmlichen Modelltests für vorgegebene Gruppen (die zur Aufteilung stetiger Variablen üblicherweise den Median als Bruchpunkt verwenden) ist, dass auch Gruppen entdeckt werden können, die sich aus anderen Bruchpunkten oder mehreren Kovariablen ergeben (z.B. Männer über 30 Jahre vs. Rest der Bevölkerung). Die zentrale methodische Herausforderung bei der rekursiven Partitionierung stellt dabei die Berücksichtigung der optimalen Auswahl des Bruchpunktes bei Kovariablen mit mehr als zwei Ausprägungen dar: Durch das Ausprobieren aller möglichen Bruchpunkte würde eine Testentscheidung anhand der herkömmlichen statistischen Verteilung zu einer Inflation des alpha-Niveaus führen. Dies muss bei der Konstruktion des Algorithmus z.B. durch die Herleitung der "richtigen" (asymptotischen oder exakten) Verteilung der optimal selektierten Statistik berücksichtigt werden. Der Vortrag zeigt die Problematik der optimal-selektierten Statistiken am Beispiel der modellbasierten rekursiven Partitionierung von Rasch-Modellen sowie am einfacheren Beispiel des chi-Quadrat-Tests auf. Aus diesen Beispielen wird klar, dass es sich hierbei um ein grundlegendes statistisches Problem handelt, das nicht nur in der rekursiven Partitionierung auftritt (und deshalb in der computationalen Statistik und Biometrie einen eigenen Forschungsschwerpunkt darstellt).

Mi., 21.09., Symposium: Methodische Innovationen in der Testdiagnostik, 15.00-15.20 Uhr, Hörsaal M3/232N (H)

Auswirkungen unterschiedlich balancierter Testheftdesigns auf Effizienz und Erwartungstreue von Parameterschätzern am Beispiel von PISA 2006

Andreas Frey

Friedrich-Schiller-Universität Jena
Institut für Erziehungswissenschaft
Professur Empirische Methoden der erziehungswissenschaftlichen Forschung
Am Planetarium 4
07737 Jena

andreas.frey
@uni-jena.de

Bei groß angelegten Schulleistungsstudien wie PISA, IGLU oder TIMSS werden sehr große Itempools verwendet. Diese enthalten zu viele Items um sie den getesteten Jugendlichen in Gänze vorzulegen, weshalb Testhefte mit jeweils einem Teil der Items zum Einsatz kommen. Die Zuordnung von Items zu Testheften erfolgt mit Testheftdesigns, häufig in Form balancierter unvollständiger Blockdesigns. Das bei PISA ab dem Jahr 2003 genutzte Testheftdesign umfasst 13 Testhefte mit jeweils 4 Positionen. Bislang liegt kaum Forschung zu den durch Testheftdesigns in der Praxis vermiedenen Verzerrungen bei der Schätzung interessierender Statistiken vor. Dies ist verwunderlich, da die Konstruktion von Testheftdesigns anspruchsvoll ist und bei der Umsetzung im Feld einen hohen Aufwand bedingt. Sollten die Vorteile sehr gering ausfallen, könnten einfacher zu handhabende Testheftdesigns Verwendung finden. Auf Basis der Daten des Vergleichs der Länder Deutschlands bei PISA 2006 (N = 36388) wurde deshalb der Frage nachgegangen, wie sich die Zerstörung der balancierten Struktur des Testheftdesigns auf die berichteten Ergebnisse auswirkt. Es wurden drei Testheftdesigns auf Länderebene bezüglich Modellpassung, Reliabilität, Mittelwerten, Standardabweichungen, Geschlechtsunterschieden und prozentualen Anteilen auf Kompetenzstufen verglichen. Betrachtet wurden das Originaldesign von PISA 2006, ein Design bei dem durch Entfernung von zwei Testheften die balancierte Designstruktur systematisch zerstört wurde und ein Design bei dem 2/13 der Antworten per Zufall gestrichen wurden. Es zeigen sich keine auffälligen Unterschiede der drei Designs hinsichtlich Modellpassung und Reliabilität. Beim systematisch gekürzten Design sind relevante Verschiebungen von Mittelwerten, der Größe von Geschlechtsunterschieden und prozentualen Anteilen auf Kompetenzstufen zu beobachten. Die Verwendung des zufällig gekürzten Designs führt erwartungsgemäß lediglich zur Erhöhung der Standardfehler der betrachteten Statistiken. Systematische Abweichungen von einer balancierten Designstruktur können somit zu relevanten Ergebnisverzerrungen führen, die nicht anhand der Modellpassung oder der Reliabilität identifiziert werden können.

Mi., 21.09., Modellierung fehlender Werte, 14.00-14.20 Uhr, Kapelle M3/126N (K)

Wie wirkt sich die Verrechnung von ausgelassenen Items auf den Item Fit im Rasch Modell aus?

Christine Hohensinn
Klaus D. Kubinger

Arbeitsbereich Psychologische Diagnostik
Fakultät für Psychologie
Universität Wien
Liebiggasse 5
1010 Wien

christine.hohensinn@univie.ac.at

klaus.kubinger@univie.ac.at

Bei der Entwicklung von Leistungstests stellen missing values durch das Auslassen von Items von der Testperson eine besondere Schwierigkeit dar. Sobald Speed-Effekte bei der Testbearbeitung ausgeschlossen werden können, bleiben die Gründe für die Nichtbearbeitung von Items zumeist unbekannt. Eine adäquate Verrechnung dieser Items ist dann schwierig wobei in der Praxis diese ausgelassenen Items meist als nicht gelöst oder aber als nicht vorgegeben verrechnet werden.
Im Zuge der Testkonstruktion werden Items mit niedrigem Item-Fit (bei vorliegenden niedrigem Gesamtmodell-Fit) aus dem Itempool ausgeschlossen und für die Letztversion des Tests nicht verwendet. Um die Auswirkung von den in der Praxis am häufigsten eingesetzten Verrechnungsregeln für missing values auf Item Fit und Modellgeltung im Rasch Modell zu untersuchen, wurde eine Simulationsstudie durchgeführt. Ausgehend von Rasch Modell konformen Daten wurde jeweils ein unterschiedlicher Anteil an missing values in den Datensatz gemäß eines MNAR-Prozesses (Rubin, 1976) implementiert. Pro Simulationsscenario wurden 1000 Replikationen durchgeführt. Die Auswirkungen auf den Item Fit sowie die Modellgültigkeit des Rasch Modells wurden jeweils für die Verrechnungsmodi nicht gelöst sowie nicht vorgegeben analysiert.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Nicht alle Hilfsvariablen für fehlende Werte sind hilfreich

Felix Thoemmes
Norman Rose

Universität Tübingen, Institut für Erziehungswissenschaft
Europastr. 6, 72072, Tübingen

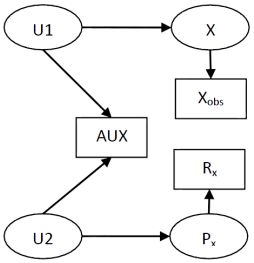


Abbildung 1

Praktisch alle empirischen Studien in den Sozialwissenschaften sind mit dem Problem fehlender Werte konfrontiert, die zunehmend mit modernen Verfahren, wie der multiplen Imputation (MI) und Full Information Maximum Likelihood (FIML) Schätzung behandelt werden. Eine Annahme dieser Verfahren ist „Missing at Random (MAR)“ (Rubin, 1976). MAR besagt, dass das Auftreten fehlender Werte bedingt stochastisch unabhängig ist von den fehlenden Variablen selbst, gegeben der beobachtbaren Variablen. Vielfach wird argumentiert das die Gültigkeit von MAR umso wahrscheinlicher ist, je mehr beobachtete Kovariaten (Hilfsvariablen) zur FIML Schätzung oder für ein Imputationsmodell verwendet werden. Weit verbreitet ist die inklusive Strategie (Collins, Schafer, & Kam, 2001) die davon ausgeht, dass irrelevante Kovariaten keine Verfälschungen der Parameterschätzungen mit sich bringen. In diesem Beitrag wird argumentiert dass es durchaus eine Klasse von Variablen gibt die nicht als Hilfsvariablen eingesetzt werden sollten, weil sie sowohl in MI und FIML die Schätzung fehlender Werte systematisch verzerren. Es handelt sich hierbei um Variablen die zwar keinen direkten Einfluss auf die Wahrscheinlichkeit fehlender Werte oder der abhängigen Variablen des Zielmodells haben, aber die eine indirekten Effekt über unbeobachtete Variablen haben (siehe Abbildung 1). Solche Variablen werden auch als Kollider bezeichnet (Pearl, 2000). Sowohl analytisch als auch durch Simulationsstudien wird in diesem Beitrag der Einfluss solcher Variablen auf die Schätzung fehlender Werte dargestellt. Die Ergebnisse zeigen, dass die inklusive Strategie Verfälschungen aufgrund fehlender Werte potentiell sogar verstärken kann anstatt zu verringern. Für die Anwendung implizieren die Ergebnisse, dass die theoriegeleiteten Auswahl von Hilfsvariablen zur FIML Schätzungen und MI notwendig sein kann. Dies bedeutet dass Forscher versuchen müssen den Einfluss von Hilfsvariablen präziser zu spezifizieren um adäquate Variablen zu selektieren.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330-351.

Pearl, J. (2000). Causality: models, reasoning, and inference. Cambridge University Press.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581.

Mi., 21.09., Modellierung fehlender Werte, 14.40-15.00 Uhr, Kapelle M3/126N (K)

Äquivalenz der Testhefte von PISA 2000 und PISA 2009

Eunike Wetzel
Claus H. Carstensen

Otto-Friedrich-Universität Bamberg, Wilhelmsplatz 3, 96047 Bamberg

eunike.wetzel@uni-bamberg.de

claus.carstensen@uni-bamberg.de

Seit mit der vierten PISA-Studie (2009) jeder der Kompetenzbereiche (Lesen, Mathematik, Naturwissenschaften) mindestens einmal der Schwerpunkt der Erhebung war, tritt zunehmend die Analyse von Trends in den Vordergrund. D.h. das Ziel von PISA ist es nicht mehr nur Ländervergleiche anzustellen, sondern auch zu untersuchen, wie sich die Kompetenzen innerhalb der teilnehmenden Länder über die Jahre verändern. (OECD, 2009) Eine Voraussetzung für die Durchführung von Trendanalysen ist die Messinvarianz der verwendeten Instrumente über Studien hinweg. Diese soll in diesem Beitrag beispielhaft für die Instrumente von PISA 2000 und PISA 2009 überprüft werden.
Anhand einer Zusatzstichprobe von 59 Gymnasien (N=4598), die an PISA 2009 teilnahmen, wurde untersucht, ob der Lesetest von PISA 2009 und der Lesetest von PISA 2000 Unterschiede aufweisen und welche Auswirkungen diese Unterschiede ggf. haben. Dazu werden die Schwierigkeiten der link Items in den Testheften von PISA 2000 und PISA 2009 mithilfe derselben Schülerinnen und Schüler verglichen. Weiterhin wird untersucht, ob sich die Ergebnisse zwischen den Jahren 2000 und 2009 für dieselben Testhefte unterscheiden und inwieweit die Testhefte in den beiden Erhebungen gleich funktionieren. Zur Beantwortung dieser Fragestellungen werden verschiedene Gleichheitsannahmen, die zum Linking der beiden Instrumente benötigt werden, anhand von restringierten IRT- Modellen (in ConQuest) überprüft.

OECD (2009). PISA 2009 assessment framework – key competencies in reading, mathematics, and science. Paris: OECD Publications.

Mi., 21.09., Modellierung fehlender Werte, 15.00-15.20 Uhr, Kapelle M3/126N (K)

Schätzfehler durch heterogene Varianzen bei Analysen mit Plausible Values

Alexander Roppelt
Christiane Penk

Humboldt-Universität zu Berlin
Institut zur Qualitätsentwicklung im Bildungswesen (IQB)
Unter den Linden 6,
10099 Berlin

Alexander.Roppelt@iqb.hu-berlin.de

In den letzten Jahren hat es sich speziell im Bildungsmonitoring etabliert, latente Fähigkeitsvariablen durch die Ziehung von multiplen Imputationen (sog. Plausible Values) der weiteren Analyse durch Standardsoftware zugänglich zu machen. So stellt etwa PISA Datensätze für Sekundäranalysen zur Verfügung, in welchen für jeden Teilnehmer und jede Fähigkeitsdimension fünf Plausible Values gezogen wurden. Diese Technik soll es ermöglichen, varianzbasierte Analysen wie beispielsweise Regressionen von der Verzerrung durch Messfehler zu befreien (vgl. Wu, 2005). Grundlage für die Ziehung der Plausible Values ist die Annahme eines Modells, worin die Dichte der latenten Fähigkeit $g(\theta)$ in der Population einer verallgemeinerten Normalverteilung folgt: $g(\theta) \sim N(b_0+b_1y_1+b_2y_2+\dots\ , \sigma^2)$. Dabei stehen die y_i für sogenannte Hintergrundvariablen, mit welchen sich die Individuen genauer beschreiben lassen. Im Falle einer einzelnen dichotomen Hintergrundvariable y_1 mit Werten 0 und 1 wäre $g(\theta)$ eine Mischverteilung aus zwei Normalverteilungen, welche um b_1 gegeneinander verschoben sind, aber die gleiche Varianz σ^2 besitzen. Generell setzt das Modell voraus, dass die Varianz für alle Ausprägungen der y_i konstant ist (Homoskedastizität). Während gezeigt werden konnte, dass das Modell recht robust gegenüber Abweichungen von der vorausgesetzten Normalform der Verteilung ist (Thomas, 2000; Walter, 2005), erweist es sich als weniger stabil gegenüber Verletzungen der Homoskedastizität (Thomas, 2000). Dies ist insofern problematisch für die Praxis der empirischen Bildungsforschung, als dass ungleiche Varianzen von Subpopulationen wie etwa in den unterschiedlichen Schulformen oder Bundesländern offenkundig sind.
In diesem Vortrag wird eine Simulationsstudie präsentiert, in welcher die Auswirkungen von heterogenen Varianzen genauer untersucht werden. Dabei wird neben den Abweichungen der geschätzten Mittelwerte und Standardabweichungen von den jeweils wahren Werten auch die Besetzung von Kompetenzstufen als Kriterium herangezogen. Auf diese Weise kann die praktische Bedeutsamkeit möglicher Verzerrungen für die Interpretation von Ergebnissen des Bildungsmonitorings besser beurteilt werden.

Thomas, N. (2000). Assessing model sensitivity of the imputation methods used in the NAEP. Journal of Educational and Behavioral Statistics, 25, 351–371.
Walter, O. (2005). Kompetenzmessung in den PISA-Studien. Simulationen zur Schätzung von Verteilungsparametern und Reliabilitäten. Lengerich: Pabst Science Publishers.
Wu, M. (2005). The role of plausible values in large-scale surveys. Studies in Educational Evaluation, 31, 114–128.

Penalisierte Spline Regression für Item Response Modelle

Otto B. Walter

Universität Bielefeld
Psychologische Methodenlehre und Qualitätssicherung
Universitätsstraße 25
33615 Bielefeld

Penalisierte Splines bieten eine Reihe von günstigen Eigenschaften, die für eine nicht-parametrische Beschreibung von Item Response Modellen herangezogen werden können. Vorteilhaft an einem solchen Ansatz ist insbesondere, dass die oft strengen Voraussetzungen parametrischer Item Response Modelle in flexibler Weise abgeschwächt werden können, wobei aber dennoch Personen- und Itemeigenschaften präzise abgebildet werden können. Der Vortrag stellt die formalen Besonderheiten des Ansatzes anhand von zwei Datensätzen aus der Leistungsdiagnostik und Persönlichkeitspsychologie vor und diskutiert Anwendungsmöglichkeiten für Skalen- und Testkonstruktion.

Eine Alternative zu nominalen Antwortformaten und deren Analyse: Die Dirichletverteilung und Dirichletregressionsmodelle in R mit DirichletReg

Marco Johannes Maier

Institute for Statistics and Mathematics
Wirtschaftsuniversität Wien
Augasse 2–6
1090 Wien

marco.maier@wu.ac.at

Kategoriale Daten und nominale Variablen im Speziellen sind ein fester und wichtiger Bestandteil in der sozialwissenschaftlichen Statistik und somit auch in der Psychologie. Die Beantwortungsmodalität eines nominalen Items ist jedem wohlbekannt – man erhält eine Anzahl disjunkter Antwortkategorien aus denen man sich für eine entscheiden muss (Multiple-Choice Formate bei denen mehrere Antwortmöglichkeiten zulässig sind stellen einen Spezialfall dar) wodurch man theoretisch gesehen eine multinomiale Verteilung erzeugt. Methodisch betrachtet analysiert man solche Daten (als abhängige Variable) demnach in einem generalisierten linearen Modell mit einer multinomialen Verteilung und einem log-Link für die linearen Prädiktoren.

Sind die Versuchspersonen ‚eindeutig' für eine der Kategorien ist dieser Ansatz angemessen – was passiert jedoch wenn sie mit mehreren Alternativen liebäugeln, aber dann nur eine wählen dürfen? Kann man solche Phänomene ignorieren oder führen sie zu Verfälschungen?

Eine Alternative findet man in der Dirichletverteilung, die als ‚kontinuierliches Pendant' zur multinomialen Verteilung gesehen werden kann und negativen Eigenschaften wie Overdispersion oder Heteroskedastizität entgegenwirkt. Des Weiteren können Versuchspersonen mehrere Kategorien mit unterschiedlichen Gewichten versehen, d.h. man erhält ein Maximum an Information und kann Verzerrungen entgegenwirken.

Im Gegensatz zu konventionellen Methoden tun sich hier viele unterschiedliche und vielversprechende Antwortmodalitäten auf, die sich sowohl mit Paper & Pencil als auch in den derzeit sehr populäreren Onlinebefragungen realisieren lassen.

Zuletzt wird die entsprechende Methodik zur Analyse solcher Daten vorgestellt, die sich Dirichletregression nennt und in der Statistiksoftware R im Paket ‚DirichletReg' (Maier, in Vorbereitung) implementiert ist.

Maier, M.J. (in Vorbereitung). DirichletReg: Dirichlet Regression for Compositional Data in R.

Eine Methode zur Parameterschätzung im Rasch-Modell bei fehlenden Werten

Jörg Heine
Christian Tarnai
Florian Hartmann

Universität der Bundeswehr München
Werner-Heisenberg-Weg 39
85577 Neubiberg

joerg-henrik.heine@unibw.de

christian.tarnai@unibw.de

Florian.Hartmann@unibw.de

Fehlende Werte sind bei sozialwissenschaftlichen Untersuchungen ein häufiges Problem. Insbesondere wenn eine Skalierung nach probabilistischen Testmodellen angestrebt wird, setzen die meisten Schätzalgorithmen, wie z.B. CML-Schätzung, eine vollständige Datenmatrix voraus. Sollen unvollständig vorliegende Datensätze ausgewertet werden, ergeben sich dann im Prinzip drei verschiedene Möglichkeiten um eine „vollständige" Datenmatrix zu erzeugen: „listweise deletion", „pairwise deletion" oder die Vervollständigung des Datensatzes mittels geeigneter Imputationsmethoden. Während die ersten beiden Möglichkeiten meist zu einer erheblichen Reduzierung der Stichprobenumfänge führen, setzen die meisten Imputationsmethoden a-priori Annahmen über den Daten-Ausfallmechanismus voraus, welche allerdings meist nicht überprüft werden können.
Die vorliegende Arbeit dokumentiert die Entwicklung eines R-Moduls, welches die explizite Itemparameterberechnung nach der Methode des paarweisen Itemvergleichs vornimmt (Choppin 1968, 1985; Zwinderman, 1995). Die Itemparameter können dabei sowohl für Items mit dichotomen Antwortformat, als auch für Items mit mehrstufigen Antwortformaten nach dem Partial Credit Modell (Masters, 1982) explizit berechnet werden (Garner & Engelhard 2009). Diese Methode hat den Vorteil, dass sie sich auch direkt auf unvollständige empirische Daten anwenden lässt. Die anschließende Schätzung der Personenparameter θ wird auf Basis der explizit berechneten Itemparameter für jede theoretisch mögliche Rohwertgruppe mittels ML-Schätzung nach der Newton-Raphson Methode vorgenommen.
Es können dabei auch (theoretische) Personenparameter für zunächst nicht empirisch beobachtete Rohwertgruppen geschätzt werden. Die Ergebnisse der Parameterschätzungen nach diesem, oben beschriebenen, Vorgehen werden anhand eines empirischen Datensatzes mit den Ergebnissen aus der CML-Schätzung mit dem Programm WinMira verglichen. Zusätzlich werden unterschiedliche Anteile von fehlenden Werten im ursprünglichen Datensatz simuliert. Abschließend soll diskutiert werden, auf welche Art und Weise die Modellgeltung anhand informationstheoretischer Kriterien (AIC, BIC, CAIC) oder globaler Modellgeltungstest für Datensätze mit fehlenden Werten vorgenommen werden kann.

Choppin, B. (1968). Item Bank using Sample-free Calibration. Nature, 219(5156), 870-872.
Choppin, B. (1985). A fully conditional estimation procedure for Rasch model parameters. Evaluation in Education, 9, 29-42.
Garner, M., & Engelhard, G., Jr. (2009). Using paired comparison matrices to estimate parameters of the partial credit Rasch measurement model for rater-mediated assessments. Journal of Applied Measurement, 10, 30-41.
Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174.
Zwinderman, A. H. (1995). Pairwise Parameter Estimation in Rasch Models. Applied Psychological Measurement, 19, 369-375.

GTT: Eine Software zur Anpassung logistischer Testmodelle

Christof Schuster
Laurens Berthold

Fachbereich Psychologie
und Sportwissenschaft
Universität Gießen
Otto-Behaghel-Str. 10
35394 Gießen

Christof.Schuster
@psychol.uni-giessen.de

GTT ist eine Software zur Analyse dichotomer Fragebogendaten auf der Grundlage logistischer Testmodelle. Als Modelle werden von GTT das 1-, 2- und 3-PL-Modell zur Verfügung gestellt. Die Anpassung des 2- und 3-PL-Modells erfolgt mit dem Marginal-Maximum-Likelihood-Ansatz. Das 1-PL-Modell lässt sich sowohl mit dem Marginal- als auch mit dem Conditional-Maximum-Likelihood-Ansatz berechnen. Auf der Grundlage von geschätzten Itemparametern lassen sich verschiedene Personenschätzer sowie Itemfit-Statistiken bestimmen. Des Weiteren können Grafiken der Itemcharakteristiken sowie der Testcharakteristik angefertigt werden. GTT wurde im Zusammenhang mit der Ausbildung von Studierenden in den Bereichen Testtheorie und Testkonstruktion an der Universität Gießen entwickelt. Dabei wurde angestrebt, die praktische Verwendung des Programms so einfach wie möglich zu gestalten.

Mi., 21.09., Anwendungsorientierte methodische Forschung I, 15.50-16.10 Uhr, Kapelle M3/126N (K)

Towards a measure of audiovisual integration efficiency

Hans Colonius

Institut für Psychologie
Carl von Ossietzky Universität Oldenburg
Ammerländer Heerstraße 114-118
26129 Oldenburg

hans.colonius@uni-oldenburg.de

Audiovisual (AV) integration refers to the process of combining information that has been extracted from the visual and auditory sensory channels. Over the last five decades, numerous studies have documented the superiority of AV speech recognition over hearing alone for nonsense syllables, words, and sentences (Calvert et al., 2004). Braida (1991) emphasized the importance of distinguishing between the ability to extract auditory and visual information, on the one hand, from an individual's capability of integrating this information from different modalities, on the other. For example, low performance of hearing-impaired listeners may not only be a consequence of a lower rate of extracting auditory information but also of a reduced efficiency in integrating it with the visual input. Similarly, it has recently been demonstrated that cochlear implant patients present higher integration efficiency in word recognition tasks when compared with normally hearing subjects, presumably due to a greater capacity to integrate visual information with the distorted speech signal (Rouger et al. 2007).

Up to now, there is no consensus about how to define a quantitative measure of integration efficiency (Grant, 2002). Here, we suggest an approach based on the theory of *Fechnerian Scaling* (Dzhafarov & Colonius, 2006). Fechnerian Scaling is an application of *dissimilarity cumulation theory* (Dzhafarov & Colonius, 2007) and deals with the computation of subjective (Fechnerian) distances among stimuli from their pairwise discrimination probabilities. The proposed measure of integration efficiency is based on relating Fechnerian distances computed from bimodal discrimination probabilities to Fechnerian distances computed from unimodal (visual and auditory) discrimination probabilities in a principled manner.

Braida, L. D.(1991). Crossmodal integration in the identification of consonant segments. Quartely Journal of Experimental Psychology, 43, 647-677.

Calvert, G., Spence, C., & Stein, B. E.(eds.) (2004). Handbook of multisensory processes. Cambridge, MA: MIT Press.

Dzhafarov, E.N., & Colonius, H. (2006). Reconstructing distances among objects from their discriminability. Psychometrika, 71, 365-386.

Dzhafarov, E.N., & Colonius, H. (2007). Dissimilarity cumulation theory and subjective metrics. Journal of Mathematical Psychology, 51(5), 290-304.

Rouger, J., Lagleyre, S., Fraysse,B . Deneve, S, Deguine, O., & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. Proceedings of the National Academy of Sciences, 104(17), 7295-7300.

Mi., 21.09., Anwendungsorientierte methodische Forschung I, 16.10-16.30 Uhr, Kapelle M3/126N (K)

Itemselektion in komplexen Zusammenhängen

Winfried Zinn

Hochschule für Gesundheit und Sport
Vulkanstraße 1
10365 Berlin

Winfried.Zinn@my-campus-berlin.com

Die Vergütung von Psychiatriepatienten erfolgt aktuell über tagesgleiche Pauschalen. Aufgrund politischer Entscheidungen soll ein leistungsorientiertes Vergütungssystem entwickelt werden. Ca. 80% aller Kosten in der Psychiatrie entstehen durch Personaleinsatz. Die durchgeführte Studie soll Hinweise geben, welche patientengebundene Indikatoren Einfluss auf den patientenspezifischen Zeitverbrauch haben. In einem Psychiatrischen Krankenhaus wurde dafür auf zwei Akutstationen mittels mobiler Erfassungsgeräte über einen Zeitraum von 5 Wochen Daten erfasst (wer hat bei welchem Patienten welche Tätigkeit wie lange durchgeführt?) In der Studie wurden bei 109 Patienten an 659 Tagen Tätigkeiten erfasst. Dabei wurden insgesamt 4343 Stunden an Tätigkeiten mit insgesamt 8321 Einzelaktionen durch Pflegekräfte und Ärzte aufgezeichnet. Darüber hinaus erfolgte retrospektiv eine Erfassung von tagesbezogenen Patientenmerkmalen, die teilweise zu unterschiedlichen Scores zusammengefasst wurden. Die Aufgabenstellung in dieser Studie stellte besondere Anforderungen an die statistischen Auswertverfahren. Es gab 266 potentielle Einflussfaktoren auf die Zielvariable (der direkt am Patienten zugeordneten Zeit). Als erster Schritt war deshalb eine Itemselektion notwendig, die diese Rahmenbedingungen bewältigen kann. Die Wahl fiel auf die Verfahren random forrest und Boosting. Diese Verfahren erlauben es, in einer nachvollziehbaren Weise Einflussfaktoren zu identifizieren, die substantiellen Einfluss auf die Zielvariable haben. Die Auswertung erfolgte mittels R und es wurden die Prozeduren Cforest (Strobl, Malley, & Tutz, 2009) (Hastie, Tibshirani, & Friedman, 2008), GLM und GBM (Hastie, Tibshirani, & Friedman, 2008) (Bühlmann & Hothorn, 2009) verwendet.

Bühlmann, P., & Hothorn, T. (2009). Twin Boosting: improved feature selection and prediction.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). The elements of statistical learning. Kap 10 & 15

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning.

Mi., 21.09., Anwendungsorientierte methodische Forschung I, 16.30-16.50 Uhr, Kapelle M3/126N (K)

Mittelwerte verstehen! Rezeption von Ergebnisberichten im Hochschulsektor

Anja Vetterlein
Erik Sengewald

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl für Methodenlehre und Evaluationsforschung
Universitätsprojekt Lehrevaluation
Am Steiger 3, Haus 1
07743 Jena

anja.vetterlein
@uni-jena.de

erik.sengewald
@uni-jena.de

Das Universitätsprojekt Lehrevaluation führt universitätsweit sowohl die Lehrveranstaltungsevaluation (Instrument PELVE; Born, Loßnitzer, & Schmidt, 2006) als auch Studiengangsbefragungen (Instrument Zwischenbilanz) an der FSU Jena durch. Pro Semester werden mehr als 500 Ergebnisberichte von Lehrveranstaltungsevaluationen an Dozenten versandt. Zusätzlich entstehen pro Semester etwa zehn umfangreiche Ergebnisberichte zu den Studiengangsbefragungen, die an fachverantwortliche Studiengangsvertreter gesendet werden. Nach Marsh (2007) ist eine Funktion der Lehrevaluation das (diagnostische) Feedback und soll der Verbesserung der Lehre dienen. Dabei reicht der bloße Empfang nicht aus, sondern die (statistischen) Ergebnisse müssen auch verstanden werden, um die richtigen Maßnahmen/Aktionen abzuleiten.
Die systematische Erforschung der Rezeption von Rückmeldungen begann mit dem Aufschwung in der Bildungsforschung. Unter anderem Müller (2010) und Schneewind (2007) untersuchten die Rezeption von Vergleichsarbeiten in der Bildungsforschung und bezogen sich dabei auf das Rahmenmodell zur Rezeption von Rückmeldungen von Helmke und Hosenfeld (2005).
Die Frage wie Akteure aus dem Hochschulsektor (Studierende, Lehrende) statistische Ergebnisse aus Lehrevaluationen verstehen und daraus Interpretationen und Handlungen ableiten, steht im Zentrum dieses Beitrags. In drei empirischen Studien wurde untersucht (a) welche Gestaltungsmerkmale in der Ergebnisdarstellung von Lehrenden vs. Studierenden präferiert werden, (b) ob die verbale Präsentation der Evaluationsergebnisse durch Experten effektiver ist und (c) wie Anschlussmaßnehmen nach der Evaluation von den Studierenden wahrgenommen werden. Die Studienergebnisse sind differenziert. Je nach Perspektive (Lehrender vs. Studierender) sind unterschiedliche Details bei der Wahrnehmung der Ergebnisse von Bedeutung.

Born, S., Loßnitzer, T., & Schmidt, B. (2006). Lehrveranstaltungsevaluation an der Friedrich-Schiller-Universität Jena – Eine Analyse der Dimensionalität der eingesetzten Fragebögen. In B. Krause & P. Metzler (eds.), Empirische Evaluationsmethoden (10, pp. 99-116). Berlin: ZeE Verlag.

Helmke, A., & Hosenfeld, I. (2005). Standardbezogene Unterrichtsevaluation. In: Brägger, G. Bucher B. und Landwehr N. (eds.), Schlüsselfragen zur externen Schulevaluation. Bern: h.e.p.-Verlag, 127-151.

Marsh, H.W. (2007). Students' Evaluations of University Teaching: Dimesnsionality, Reliabilty, Validity, Potential Biases and Usefulness. In: R.P. Perry and J.C. Smart (eds.), The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective, 319-383.

Mi., 21.09., Anwendungsorientierte methodische Forschung I, 16.50-17.10 Uhr, Kapelle M3/126N (K)

Eine neue Methode zur Schätzung der Varianzüberschneidung zwischen der Kurz- und der Langform eines psychologischen Tests

Nils Pfeiffer[1]
Dirk Hagemann[2]
Matthias Backenstrass[3]

[1]Schön Klinik Roseneck, Am Roseneck 6, 83209 Prien am Chiemsee

npfeiffer@schoen-kliniken.de

[2]Psychologisches Institut Universität Heidelberg

[3]Klinik für Allgemeine Psychiatrie, Zentrum für Psychosoziale Medizin Universitätsklinikum Heidelberg

G. T. Smith, D. M. McCarthy und K. G. Anderson (2000) stellten Richtlinien zur Konstruktion und Evaluation von Kurzversionen psychologischer Tests vor, um die methodischen Standards in diesem Bereich anzuheben. Entsprechend einer ihrer Richtlinien sollen die Entwickler von Kurzversionen zeigen, dass die Kurzversion die gleiche Variable erfasst wie die zugehörige Langversion. Bisher gab es kein methodisch stimmiges Verfahren, um die Varianzüberschneidung zwischen Kurz- und Langversion zu schätzen. Die neue VO Methode wird diesem Anspruch gerecht (Pfeiffer et al., 2011). Entsprechend dieses Verfahrens werden Probanden auf vier Gruppen randomisiert, die mit zeitlichem Abstand die Kurz- und die Langversion in unterschiedlichen Kombinationen bearbeiten (kurz-kurz, lang-lang, lang-kurz, kurz-lang). Unter Einbezug der Minderungskorrektur von Spearman (1904) erlauben die aus den vier Testkombinationen resultierenden vier Korrelationen die erwartungstreue Schätzung der Varianzüberschneidung zwischen Lang- und Kurzversion.

Pfeiffer, N., Hagemann, D., & Backenstrass, M. (2001). A new method for estimating the variance overlap between the short and the long form of a psychological test. Educational and Psychological Measurement, 71, 380-388.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. Psychological Assessment, 12, 102-111.

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 3, 271-295.

TAGUNGSBEITRÄGE

Donnerstag

Do., 22.09., Symposium: Items, responses, and theory, 9.00-11.00 Uhr, Hörsaal M3/232N (H)

Symposium: Items, responses, and theory

Steffi Pohl[1]
Jan Marten Ihme[2]
[1]Otto-Friedrich-Universität Bamberg
NEPS
96045 Bamberg

steffi.pohl@uni-bamberg.de

[2]Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) an der Universität Kiel
Olshausenstraße 62
24118 Kiel

ihme@ipn.uni-kiel.de

Die Item-Response-Theorie stellt Modelle bereit, mit denen die Antworten von Personen auf Items betrachtet werden können. Dabei werden die Fähigkeiten von Personen und die Schwierigkeiten von Items auf einer gemeinsamen Skala angeordnet. Somit können Modelle formuliert werden, die es erlauben Eigenschaften der Personen, Eigenschaften der Items oder auch beides zugleich zu erklären/modellieren.

In diesem Symposium werden Beiträge zu Eigenschaften von Items und Personen sowie zum Umgang mit Responses präsentiert. Dabei wird auf Theorien bei der Modellierung zurückgegriffen. Hartig untersucht interindividuelle Unterschiede in Itempositionseffekten und modelliert damit sowohl Eigenschaften von Personen als auch von Items. Ihme präsentiert ein theoriegeleitetes Vorgehen zur Untersuchung von DIF und modelliert damit ebenfalls Eigenschaften von Items für verschiedene Personengruppen. Rose und von Davier sowie Pohl, Gräfe und Hardt fokussieren auf die Modellierung fehlender Antworten (Responses) und beschäftigen sich mit der Dimensionalität und Bedeutung der modellierten latenten (Missing-) Variable (Rose & von Davier) oder mit der Ignorability und Modellierung der fehlenden Werte (Pohl, Gräfe & Hardt). Item-Response-Modellierungen ermöglichen auch die Betrachtung von Trends, auch wenn in verschiedenen Erhebungswellen Responses auf unterschiedliche Items gegeben wurden. Im Beitrag von Robitzsch wird die Berechnung von Standardfehlern aufgrund von Item-Sampling für Ländermittelwertsvergleiche sowie Trendschätzungen vorgeschlagen, während sich der Beitrag von Carstensen, Lankes und Steffinsky mit Modellen für Linking in Trendanalysen beschäftigt.

Beiträge:

1. Hartig, Johannnes & Buchholz, Janine: Analyse von Itempositionseffekten und interindividuellen Unterschieden in diesen Effekten in einem Mehrebenen-IRT-Modell
2. Ihme, Jan Marten: Ein anderer Blick auf DIF oder Eine Strategie für Schiffe versenken
3. Rose, Norman & von Davier, Matthias: IRT Modelle für nonignorable Missing Data bei komplexer Dimensionalität
4. Pohl, Steffi, Gräfe, Linda & Hardt, Katinka: Ignorability und Modellierung von fehlenden Werten in Kompetenztests
5. Robitzsch, Alexander: Berechnung von Standardfehlern in Large-Scale Assessments auf Grund von Item Sampling für Ländermittelwerte, originale und marginale Trendschätzungen
6. Carstensen, Claus H., Lankes, Eva-Maria & Steffinsky, Mirjam: Homogenität von Kompetenzverläufe über drei Messzeitpunkte am Beispiel der Studie zu Naturwissenschaftlichen Kompetenzentwicklung im Elementarbereich – SnaKE

Analyse von Itempositionseffekten und interindividuellen Unterschieden in diesen Effekten in einem Mehrebenen-IRT-Modell

Johannes Hartig
Janine Buchholz

DIPF - Deutsches Institut für Internationale Pädagogische Forschung
Schloßstr. 29
60486 Frankfurt/Main

hartig@dipf.de

buchholz@dipf.de

In standardisierten Tests, wie sie z. B. in großen Schulleistungsstudien zum Einsatz kommen, kann sich die Leistung der Schülerinnen und Schüler im Testverlauf aufgrund von Ermüdung oder abnehmender Motivation verringern, d. h. die Itemposition hat Effekte auf die Itemschwierigkeit. In den Standard-Modellen zur Testauswertung (z. B. dem Raschmodell) sind diese Effekte nicht enthalten, sie können jedoch durch eine Erweiterung in einem Mehrebenen-Ansatz, bei dem Antworten als geschachtelt in Individuen betrachtet werden, berücksichtigt und geschätzt werden. Zusätzlich zu den festen Effekten der Itemposition können durch die Aufnahme von Zufallseffekten in das Modell auch interindividuelle Unterschiede im Ausmaß der Itempositionseffekte geschätzt werden. Diese Zufallseffekte können als interindividuelle Unterschiede in der Ausdauer bei der Testbearbeitung interpretiert werden.

In der vorliegenden Studie werden Itempositionseffekte und damit verbundene interindividuelle Unterschiede anhand der Daten zu naturwissenschaftlichen Kompetenzen aus der internationalen PISA 2006-Studie analysiert. Untersucht werden das Ausmaß der festen Itempositionseffekte, der Umfang der diesbezüglichen interindividuellen Differenzen sowie die Korrelation zwischen dem individuellen Leistungsniveau und der individuellen Ausdauer. Insgesamt liegen Testdaten für 397.920 Schülerinnen und Schüler aus 57 Ländern vor. Die Analysen wurden für ausgewählte Länder vorgenommen, die das gesamte Spektrum des internationalen Leistungsspektrums abdecken. Konsistent über alle Länder findet sich ein kleiner Effekt der Itemposition auf die Itemschwierigkeit. In allen Ländern zeigten sich signifikante interindividuelle Differenzen in der Ausdauer, diese Effekte sind jedoch in Ländern mit niedrigem Leistungsniveau deutlich ausgeprägter. Die Korrelation zwischen individuellem Leistungsniveau und individueller Ausdauer ist in Ländern mit hohem nationalem Leistungsniveau praktisch null, in Ländern mit niedrigem nationalem Leistungsniveau hingegen negativ. Mögliche Erweiterungen des Analysemodells werden diskutiert.

Ein anderer Blick auf DIF *oder* Eine Strategie für Schiffe versenken

Jan Marten Ihme

Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) an der Universität Kiel
Olshausenstraße 62
24118 Kiel

ihme@ipn.uni-kiel.de

Ein übliches Vorgehen bei DIF-Analysen ist die sog. Item Purification, bei der aus einem Test iterativ jeweils ein oder zwei Items mit den deskriptiv höchsten DIF-Werten ausgeschlossen und die DIF-Werte neu berechnet werden (z. B. Holland & Thayer, 1988). Dies führt entsprechend zu einem bereinigten Test. Es wird jedoch nicht überprüft, inwieweit dadurch systematische oder Zufallseffekte verantwortlich sind.

Auch bei dem Spiel Schiffe versenken können durch zufällige Schüsse durchaus Schiffe des Gegners versenkt werden. Durch Kenntnis der Regeln für das Platzieren der Schiffe (z. B. Größe der Schiffe, dürfen Schiffe direkt nebeneinander platziert werden) können jedoch Strategien angewendet werden, die zufälligen Schüssen deutlich überlegen sind. Ich schlage vor, auch bei der Suche nach DIF Strategien anzuwenden, die beim Detektieren von DIF und dem Formulieren von Erklärungen dafür einem normativen Ausschlusskriterium für Items überlegen sind.

Dazu sollte eine DIF-Analyse mit einem globalen Test beginnen, mit dem festgestellt wird, ob prinzipiell DIF in den Items vorhanden ist. Dafür sind in der Literatur zahlreiche Methoden zu finden (für eine aktuelle Zusammenfassung s. z. B. Magis, Béland, Tuerlinckx & De Boeck, 2010). Soweit auf globaler Ebene DIF gefunden wurde, empfehle ich, als zweiten Schritt Hypothesen darüber aufzustellen, welche Itemeigenschaften für DIF verantwortlich sein können. In einem dritten Schritt sollten diese Hypothesen überprüft werden, indem die DIF-Werte der Items mit den Item-Eigenschaften in Verbindung gebracht werden.

Zusammenfassend empfehle ich, beim Schiffe versenken nicht nach dem Zufallsprinzip zu schießen, sondern das Wissen über die Regeln zu nutzen und mit einer Strategie die Anzahl der nötigen Schüsse zu reduzieren. Analog empfehle ich für DIF-Analysen, das Wissen über die Items auszunutzen, um DIF zu identifizieren, statt einfach Items nach einem normativen Kriterium aus dem Test zu entfernen.

Holland, P. W. & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Erlbaum.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. Behavior Research Methods, 42, 847-862.

IRT Modelle für nonignorable Missing Data bei komplexer Dimensionalität

Norman Rose[1]
Matthias von Davier[2]

[1] Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl für Methodenlehre und Evaluationsforschung
Am Steiger 3, Haus 1
07743 Jena

norman.rose@uni-jena.de

[2] Educational Testing Service, Princeton, New Jersey (USA)

Modellbasierte Verfahren zur Behandlung fehlender Werte finden auch in der IRT zunehmend Verbreitung (O'Muircheartaigh & Moustaki, 1999, Moustaki & Knott, 2000, Holman & Glas, 2005, Rose et al. 2010). Verschiedene mehrdimensionale IRT (MIRT) Modelle sind entwickelt worden, die auf der gemeinsamen Modellierung der Testitems und deren Response-Indikatoren basieren. Neben der latenten Fähigkeitsvariablen wird eine Response Propensity als zusätzliche latente Variable konstruiert deren Messmodell auf Response-Indikatoren basiert. Verschiedene MIRT Modelle sind vorgeschlagen worden, die grundsätzlich in Between-item-dimensional und within-item-dimensional Models unterschieden werden können. Die Modelle werden oft als äquivalent angesehen was jedoch nicht per se gilt. Die Entscheidung für ein bestimmtes Modell hängt von verschiedenen Faktoren ab und muss für mehrdimensionale Raschmodelle anders beantwortet werden als für zweiparametrische Modelle. Die Dimensionalität der latenten Response Propensity wird häufig nicht berücksichtigt ist aber ebenfalls relevant bei der Wahl eines geeigneten Modells. Darüber hinaus ist der Begriff der Äquivalenz von Messmodellen im Kontext modellbasierter Verfahren für fehlende Werte schwierig. Es wird erläutert hinsichtlich welcher Aspekte Modelle äquivalent zueinander sein müssen, um in gleichem Maße als geeignet zu gelten. Schließlich werden verschiedene Modelle vergleichend betrachtet und Unterschiede hinsichtlich der Bedeutung und Interpretation einzelner Modellparameter verdeutlicht.

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

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. Journal of the Royal Statistical Society, SeriesA, 163, 445-459.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitudes scales. Journal of Royal Statistic Society, 162, 177-194.

Rose, N., Davier, M. von, & Xu, X. (2010). Modeling nonignorable missing data with IRT (Research Report No. RR-10-11). Princeton, NJ: Educational Testing Service.

Do., 22.09., Symposium: Items, responses, and theory, 10.00-10.20 Uhr, Hörsaal M3/232N (H)

Ignorability & Modellierung von fehlenden Werten in Kompetenztests

Steffi Pohl
Linda Gräfe
Katinka Hardt

Otto-Friedrich-Universität Bamberg
NEPS
96045 Bamberg

steffi.pohl
@uni-bamberg.de

In Daten zur Messung von Kompetenzen mit Hilfe von Tests treten häufig fehlende Werte auf. Dabei sind vor allem fehlende Werte, die durch Überspringen von Items oder durch Nicht-Erreichen aufgrund von Zeitbegrenzung entstehen, nicht ignorierbar (nonignorable, Mislevy & Wu, 1996) und können zu einer verzerrten Schätzung der Item- und Personenparameter führen. Es gibt verschiedene Methoden zum Umgang mit fehlenden Werten, wobei vor allem modellbasierte Ansätze große Beachtung finden, da sie die Untersuchung der Abhängigkeit der fehlenden Werte von der Fähigkeit der Person ermöglichen (Glas & Pimentel, 2008; Holmann & Glas, 2005; Rose, von Davier, & Xu, 2010). In verschiedenen Studien wurden diese Methoden bereits angewendet, jedoch meist ohne dabei zwischen übersprungenen und nicht erreichten Items zu unterscheiden. Der generierende Prozess für das Auftreten fehlender Werte ist jedoch nicht unbedingt derselbe für diese beiden Arten fehlender Werte. In unserer Studie wird die Abhängigkeit der fehlenden Werte von der Fähigkeit der Person getrennt für beide Arten fehlender Werte, übersprungene sowie nicht erreichte Items, an den Kompetenztestdaten des Nationalen Bildungspanels untersucht. Dabei wird die Tendenz zu fehlenden Werten für verschiedene inhaltliche Kompetenzdomänen betrachtet und die Stabilität dieser Tendenz über die verschiedenen Domänen untersucht. Es werden verschiedene Spezifikationen modellbasierter Ansätze zum Umgang mit fehlenden Werten an den Daten getestet und die Item- und Personenparameterschätzung mit den Ergebnissen aus entsprechend simulierten Daten verglichen. Die Ergebnisse dieser Studie tragen zur Erklärung des Auftretens fehlender Werte in Kompetenztests bei und geben praktische Hinweise zur Modellierung von fehlenden Werten.

Glas, C. A. W., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. Educational and Psychological Measurement, 68, 907-922.

Holman, R., & Glas, C. A. W. (2005). Modelling nonignorable missing data mechanisms with item response theory models. British Journal of Mathematical and Statistical Psychology. 58, 1–17.

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

Rose, N., von Davier, M., & Xu, X. (2010). Modeling Nonignorable Missing Data With Item Response Theory (Research Report RR-10-11). Princeton, NJ: Educational Testing Service.

Do., 22.09., Symposium: Items, responses, and theory, 10.20-10.40 Uhr, Hörsaal M3/232N (H)

Berechnung von Standardfehlern in Large-Scale Assessments auf Grund von Item Sampling für Ländermittelwerte, originale und marginale Trendschätzungen

Alexander Robitzsch

Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens (BIFIE)
Salzburg
Zentrum für Bildungsmonitoring & Bildungsstandards
Alpenstraße 121
5020 Salzburg, Österreich

a.robitzsch@bifie.at

In Large-Scale Assessments (z. B. der internationalen Schulleistungsstudie PISA) werden mehrere Ländermittelwerte interessierenden Domänen (wie der Lesekompetenz) verglichen. Diese Studien verwenden Item-Response-Modelle (z. B. das Raschmodell) als Grundlage der Schätzung von Personenfähigkeiten. Oft wird dabei jedoch die Auswahl von Items (sog. Item Sampling) als Variabilitätsquelle für Mittelwertvergleiche ignoriert (Brennan, 2011).
Dieser Beitrag definiert einen Standardfehler für einen Ländermittelwert auf Grund von Item Sampling auf Basis der Varianz von Effekten differenziellen Itemfunktionierens (DIF-Varianz) eines Landes. Außerdem werden verschiedene Techniken zur Ermittlung von Standardfehlern für originale und marginale Trendschätzungen diskutiert (Carstensen, Prenzel & Baumert, 2008; Monseur & Berezner, 2007), die mit der Auswahl von Items zusammenhängen.
Anhand der österreichischen PISA-Daten in den Studien 2000, 2003, 2006 und 2009 werden die durch Item Sampling bedingten Standardfehler für Querschnitts- und Trendschätzungen für die Domäne der Lesekompetenz vorgestellt und deren Implikation ausgeführt.

Brennan, R. L. (2011). Generalizability theory and classical test theory. Applied Measurement in Education, 24, 1-21.

Carstensen, C. H., Prenzel, M. & Baumert, J. (2008). Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt? Zeitschrift für Erziehungswissenschaften, 10, 11-34.

Monseur, C. & Berezner, A. (2007). The computation of equating errors in international surveys in education. Journal of Applied Measurement, 8, 323-335.

Do., 22.09., Symposium: Items, responses, and theory, 10.40-11.00 Uhr, Hörsaal M3/232N (H)

Homogenität von Kompetenzverläufe über drei Messzeitpunkte am Beispiel der Studie zu Naturwissenschaftlichen Kompetenzentwicklung im Elementarbereich - SnaKE

Claus H. Carstensen[1]
Eva-Maria Lankes[2]
Mirjam Steffensky[3]

[1]Universität Bamberg

claus.carstensen@uni-bamberg.de

[2]TU München
[3]IPN Kiel

Ziel des Projektes ist es, zu untersuchen über welche naturwissenschaftlichen Kompetenzen Vorschulkinder verfügen und inwieweit eine gezielte Intervention den Aufbau dieser Kompetenzen vorbereiten und unterstützen kann. Die Fragestellungen des Projekts beziehen sich dabei auf den Aufbau von naturwissenschaftlichem Wissen sowie in Ansätzen Wissen über die Fähigkeit, dieses Wissen anzuwenden und das Interesse an naturwissenschaftlichen Sachverhalten. Die naturwissenschaftliche Kompetenz wird exemplarisch anhand des Themenbereiches "Wasser" untersucht, wobei die Teilaspekte "Schmelzen und Gefrieren", "Verdunsten/Verdampfen und Kondensieren" und "Lösen/Nicht-Lösen" betrachtet werden.

Zu insgesamt drei Messzeitpunkten, vor einer Intervention, vier Monate später nach der Intervention sowie drei Monate später, wird naturwissenschaftliche Kompetenz mit einem Test in Form eines strukturierten Interviews erfasst. Anhand der Erhebung zum ersten Messzeitpunkt konnte gezeigt werden, dass sich mit dem entwickelten Instrumenten naturwissenschaftliche Kompetenz von Fünfjährigen abbilden lässt.

Im vorgeschlagenen Beitrag wird die Analyse der Kompetenzverläufe über alle Erhebungen dargestellt. Dazu wird zunächst die Vergleichbarkeit der Aufgaben zwischen den Erhebungen untersucht. Anschließend wird die Verbindung der Testergebnisse auf eine gemeinsame Skala dargestellt. Gegeneinander geprüft werden dabei ein Linking-Modell, welches über die Verteilung der Itemschwierigkeitsparameter eine Verbindung herstellt (mean-linking; Kolen & Brennan, 2004) und ein Equating-Modell über eine gemeinsame Skalierung aller Daten (concurrent calibration; ebd.). Es zeigt sich, dass die zu jedem Messzeitpunkt Rasch-homogenen Tests sich nicht dazu eignen, einen Kompetenzverlauf auf einer allen drei Messzeitpunkt gemeinsamen Skala abzubilden. In der Diskussion werden alternative Skalierungsmethoden aufgegriffen.

Kolen, M. J. & Brennan, R. L. (2004). Test Equating, Scaling, and Linking: Methods and Practices (2nd ed.). New York: Springer.

Alternative Verwendung eines CS-C(M−1) Modells: Erfassung von Domänenspezifität durch Residualfaktoren

Tanja Könen
Franzis Preckel
Martin Brunner

Deutsches Institut für Internationale Pädagogische Forschung (DIPF)
Schloßstraße 29
60486 Frankfurt am Main

koenen@dipf.de

Im Correlated State-Correlated (Methods-Minus-1) Modell werden unterschiedliche Methoden zur Erfassung von State-Variablen als Residualfaktoren einer Referenzmethode modelliert (Geiser, 2009). Dabei sind die Residualfaktoren unter einem generellen Faktor genestet, welcher die State-Variable – gemessen mit der Referenzmethode – repräsentiert. Das Modell eignet sich über eine klassische Multitrait-Multimethod-Anwendung hinaus insbesondere dafür, unter einer längsschnittlichen Perspektive die generelle und domänenspezifische Varianz eines Konstrukts zu differenzieren. Dies wird am Beispiel des akademischen Selbstkonzeptes illustriert. 879 Schülerinnen und Schüler beantworteten an vier Messzeitpunkten in der fünften und sechsten Gymnasialklasse Items zum allgemeinen akademischen Selbstkonzept (z.B. „In den meisten Schulfächern lerne ich schnell“) und zu domänenspezifischen Selbstkonzepten in Mathematik und Deutsch (z.B. „Im Fach Mathematik/ Deutsch lerne ich schnell“). In einem State-Modell werden die domänenspezifischen Selbstkonzepte als Residualfaktoren des allgemeinen akademischen Selbstkonzeptes modelliert und beinhalten somit Varianz, die spezifisch für das jeweilige Schulfach ist. Das State-Modell besitzt einen guten globalen und lokalen Modell-Fit. Alle Faktoren erweisen sich als stabil und skalar messinvariant über die Zeit. Somit lässt sich die domänenspezifische Selbstkonzeptentwicklung unter Kontrolle der allgemeinen akademischen Selbsteinschätzung betrachten. Weitere Vorteile und Nachteile des Modells werden diskutiert.

Geiser, C. (2009). Multitrait-Multimethod-Multioccasion Modeling. München, Germany: Akademische Verlagsgemeinschaft.

Do., 22.09., Modellierung latenter Variablen, 9.20-9.40 Uhr, Kapelle M3/126N (K)

Die räumliche Dimension in der Psychometrie: Ein latentes 'Geopress-State'-Modell

Rüdiger Mutz

Professur für Sozialpsychologie und Hochschulforschung
ETH Zürich
CH-8092 Zürich

mutz@gess.ethz.ch

Mittels eines psychologischen Tests wird eine Person zuerst einmal in ihrem Verhalten in einer konkreten Situation erfasst (Steyer & Schmitt, 1990). Latent state-trait Modellen erlauben es, über Wiederholungsmessungen von Konstrukten über Situationen und Zeitpunkten das beobachtete Verhalten von situationsspezifischen Faktoren und Messfehler zu bereinigen, um tatsächlich zu den interessierenden zeitstabile Dispositionen (latent traits) zu kommen. Während bisher nur die zeitliche Dimension von Situationen gesehen wurde, soll in diesem Beitrag die räumliche Dimension einbezogen werden (Mutz & Becker, 2002). Hierbei verändert sich die Perspektive, von der Stabilität der Person (trait) auf die lokalen Stabilität einer räumliche Situation, die ausgehend von Murrays need-press-Konzept mit dem Begriff „latent geopress" bezeichnet werden soll. Sozialpsychologischer Hintergrund bildet die raumbezogene Social Impact Theorie von Latané (Latané, Nowak & Liu, 1994). Ein „latent geopress" liegt vor, wenn Personen in einem lokalen räumlichen, sprich geopositionierten, Kontext systematisch ähnliches Verhalten zeigen. Ziel dieses methodischen Beitrags ist es, ausgehend von dem Latent state-trait Modell von Steyer ein latentes Geopress-state Modell auf der Grundlage eines räumlichen Strukturgleichmodells (Congdon, 2009; Oud & Folmer, 2007) darzustellen, das es erlaubt, die verschiedenen Varianzanteile (latent geopress, latent state) zu schätzen. Eine Reanalyse von Daten zu Einstellungen von Privatwaldbesitzerinnen und –besitzer (N=949) zu Ihrem Wald in NRW, deren Stichprobe anhand eines 1 km × 1 km geopositionierten Netz (Waldpunkte) aufgenommen wurde, soll den Ansatz beispielhaft illustrieren (Mutz & Becker, 2002).

Congdon, P. (2009). A spatial structural equation model with an application to area health needs. Journal of Statistical Computation and Simulation, 80(4), 401-412.

Latané, B., Nowak, A. & Liu, J. H. (1994). Measuring emergent phenomena – dynamism, polarization, and clustering as order parameters of social systems. Behavioral Science, 39(1), 1-24.

Mutz, R. (2002). "Attitude mining" – Analyse räumlicher Muster von Waldbesitzermotiven in Nordrhein-Westfalen. Zeitschrift für Sozialpsychologie, 33(2), 101-109.

Oud, J. H. L. & Folmer, H. (2008). A structural equation approach to models with spatial dependence. Geographic Analysis, 40, 152-166.

Steyer, R. & Schmitt, M. J. (1990). Latent state-trait models in attitude research. Quality & Quantity, 24, 427-445.

Do., 22.09., Modellierung latenter Variablen, 9.40-10.00 Uhr, Kapelle M3/126N (K)

Faktorstrukturinvarianz = Faktorinvarianz: Ein weit verbreiteter Irrglaube

Moritz Heene[1]
Michael Maraun[2]

[1]Institut für Psychologie
Karl Franzens Universität Graz
Maiffredygasse 12b
A-8010 Graz

moritz.heene
@uni-graz.at

[2]Department of Psychology
Simon Fraser University
8888 University Drive
Burnaby
B.C. V5A 1S6 Burnaby
Canada

michael_maraun
@sfu.ca

Im Rahmen von konfirmatorischen Multigruppenfaktorenanalysen findet sich in der Literatur häufig die Behauptung, dass die Invarianz von Strukturparametern wie beispielsweise Faktorladungen und Fehlervarianzen zwischen Populationen Antworten auf die Frage nach der Äquivalenz bzw. Ähnlichkeit der Faktoren selbst liefert. Die Studie zeigt anhand theoretischer Ausführungen und einer empirischen Illustration auf, dass aufgrund der statistischen Definition der Zufallsvariablenäquivalenz die Multigruppenfaktorenanalyse keine hinreichende Grundlage für die Bestimmung der Faktoräquivalenz/Faktorähnlichkeit liefert. Die Ergebnisse zeigen, dass Faktoren, deren Strukturparameter über Populationen invariant sind vollkommen verschieden sein können. Weiterhin wird dargestellt, dass es sich bei der in der Literatur verbreiteten Gleichsetzung von Faktorstrukturäquivalenz mit Faktoräquivalenz um eine Verwechslung von mehrdimensionalen Verteilungsparametern mit Zufallsvariablen handelt. Es wird daher aufgezeigt, dass eine Unterscheidung zwischen Faktorstrukturäquivalenz und Faktoräquivalenz für die korrekte inhaltliche Interpretation von Ergebnissen aus Multigruppenfaktorenanalysen unabdingbar ist.

Do., 22.09., Modellierung latenter Variablen, 10.00-10.20 Uhr, Kapelle M3/126N (K)

Eine Monte-Carlo Studie zum Vergleich der Genauigkeit des ULS-, DWLS- und PIV-Schätzers bei dichotomen konfirmatorischen Faktorenanalysen

Steffen Nestler

Johannes Gutenberg-Universität Mainz
Institut für Psychologie
Lehrstuhl für Persönlichkeitspsychologie und Psychologische Diagnostik
55099, Mainz

Die Parameter (z.B. Faktorladungen, Varianzen der Faktoren) des dichotomen konfirmatorischen Faktorenanalysemodells werden üblicherweise mittels des ULS- (unweighted least squares; Muthén, 1993) oder des DWLS- Schätzverfahrens (diagonally weighted least squares; Muthén, du Toit, & Spisic, 1997) bestimmt. Bollen und Maydeu-Olivares (2007) haben kürzlich eine alternative Schätzmethode vorgeschlagen (PIV: polychoric instrumental variables), die im Gegensatz zu ULS und DWLS die Parameter nicht in einem Schritt bestimmt, sondern bei der zunächst die Faktorladungen mittels Instrumentenvariablen berechnet werden, um daraufhin die Varianzen der Faktoren und die Kovarianzen zwischen den Faktoren zu schätzen. Wir haben eine Monte-Carlo Simulationsstudie mit dem Ziel durchgeführt, die Genauigkeit der drei Verfahren bei der Schätzung der Parameter eines dichotomen Faktorenmodells miteinander zu vergleichen. Die Simulation umfasste dabei 48 Bedingungen, die durch eine Kombination von (a) vier Stichprobengrößen (100, 250, 500, 1000), (b) drei Faktorladungen (gering, moderat, hoch), (c) drei Stufen der Schwere des Verstoßes gegen die Normalverteilung (normal, moderat nicht-normal, extrem nicht-normal) und (d) ob das Modell korrekt oder falsch spezifiziert war, definiert wurden. Die Ergebnisse zeigten, dass wenn das Modell korrekt spezifiziert wurde, PIV genauso akkurate Parameterschätzungen erbrachte wie ULS und DWLS. Genauere Schätzungen durch PIV ergaben sich im Falle des falsch spezifizierten Modells. Implikationen der Ergebnisse werden diskutiert.

Bollen, K. A., & Maydeu-Olivares, A. (2007). Polychoric instrumental variable (PIV) estimator for structural equations with categorical variables. Psychometrika, 3, 309–326.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551–560.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.

Do., 22.09., Modellierung latenter Variablen, 10.20-10.40 Uhr, Kapelle M3/126N (K)

Einfluss der hierarchischen Struktur auf den Modellfit für Instrumente der Lehrevaluation

Erik Sengewald
Anja Vetterlein

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl für Methodenlehre und Evaluationsforschung
Universitätsprojekt Lehrevaluation
Am Steiger 3 Haus 1
07743 Jena

erik.sengewald@uni-jena.de

anja.vetterlein@uni-jena.de

Lehrevaluation wird an der Friedrich-Schiller-Universität bereits seit 1997 durch das Universitätsprojekt Lehrevaluation durchgeführt. Bis zum Jahre 2004 entwickelte das Projekt das Evaluationsinstrument PELVE (Prozess und Ergebnisorientierte Lehrveranstaltungsevaluation) zur mehrdimensionalen Erfassung der Lehrqualität. Durch die Anforderungen seitens der Lehrenden, sind die Bögen sehr kurz und trotzdem facettenreich um dem Nutzen eines formativen Feedbacktools gerecht zu werden. Dennoch werden diese Evaluationsinstrumente nach psychometrischen Kriterien. Eine Übersicht hierzu bietet Braun (2008). Meistens werden dafür exploratorische Faktorenanalysen verwendet (Marsh et al., 2009). Die wenigen konfirmatorischen Faktorenanalysen zeigen meist keinen akzeptablen Fit (CFI, TLI >.9; RMSEA <.05) (Marsh et al., 2009). Für den PELVE liegen konfirmatorische Analysen vor (Born, Loßnitzer, & Schmidt, 2006). Die Autoren schlagen ein fünf-Faktoren Modell mit Faktoren zweiter Ordnung vor. Bei der Überprüfung der theoretischen Dimensionen in den Instrumenten zur Lehrevaluation wurde in den vorliegenden Studien die hierarchische Datenstruktur jedoch nicht berücksichtigt. Die vorliegende Studie geht auf die hierarchische Ordnung der Daten ein und zeigt den Einfluss dieser Struktur auf die Fitstatistik im Rahmen konfirmatorischer Faktorenanalysen und exploratorischen Strukturgleichungsmodellen. Durch einen spezifischen Personencode gelingt es den Datensatz von insgesamt 129.258 Personen um Studierenden- und Dozentendopplungen zu bereinigen, sodass die Analysen auf Grundlage der bereinigten Veranstaltungsmittelwerte durchgeführt werden können. Die Ergebnisse zeigen, dass spezifizierte Modelle ohne Berücksichtigung der hierarchischen Struktur einen besseren Fit aufweisen, als Modelle unter Berücksichtigung der hierarchischen Struktur.

Born, S., Loßnitzer, T., & Schmidt, B. (2006). Lehrveranstaltungsevaluation an der Friedrich-Schiller-Universität Jena – Eine Analyse der Dimensionalität der eingesetzten Fragebögen. In B. Krause & P. Methler (Hrsg.), Empiri sche Evaluationsmethoden (10, pp. 99 – 116). Berlin: ZeE Verlag.

Braun, E. (2008). Das Berliner Evaluationsinstrument für selbsteingeschätzte studentische Kompetenzen - BEvaKomp. Göttingen: Vandenhoeck & Ruprecht unipress.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S. et al. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. Structural Equation Modeling, 16, 439-476.

Do., 22.09., Modellierung latenter Variablen, 10.40-11.00 Uhr, Kapelle M3/126N (K)

Anwendung von Latente-Variablen-Modellen zur Testgütekriterienbestimmung bei Testverfahren zur Erfassung nominaler Konstrukte

Hendryk Böhme

Abt. für Angewandte Psychologische Forschung und Entwicklung

Bundesagentur für Arbeit
Regensburger Straße 104
90478 Nürnberg

hendryk.boehme
@arbeitsagentur.de

Sehr häufig werden die mit Psychologischen Tests gemessenen Konstrukte als kontinuierliche latente Variablen angenommen und entsprechend modelliert. Dies gilt in besonderem Maße für Leistungstests. Zur Bestimmung von Testgütekriterien existieren im Falle kontinuierlicher latenter Variablen vielfältig erprobte Methoden und Analyseverfahren. Viele dieser Verfahren sind jedoch nicht anwendbar, wenn das gemessene Merkmal als nominale latente Variable konstruiert wird. Am Beispiel des (nominalen) Konstruktes Sophistiziertheit Konditionalen Schlussfolgerns werden Latente-Variablen-Modelle bzw. Analyseverfahren vorgestellt, die die Bestimmung wichtiger Testgütekriterien wie Reliabilität und Validität ermöglichen. Diese Analyseverfahren (z.B. Analyse latenter Klassen, Analyse latenter Transitionen, multinomiale logistische Regression für latente Variablen) liefern adäquate Informationen und gut interpretierbare Ergebnisse und bieten sich daher generell für die Entwicklung und Erprobung psychodiagnostischer Testverfahren zur Erfassung nominaler Konstrukte an.

Tücken des Bootstrap

Rainer W. Alexandrowicz
Clemens Draxler

Alpen-Adria-Universität Klagenfurt
Abteilung für Angewandte Psychologie und Methodenforschung
Universitätsstraße 65–67
9020 Klagenfurt
Österreich

Die Überprüfung der Gültigkeit des Rasch-Modells mittels des Likelihood-Ratio-Tests nach Andersen (1973) zählt bereits seit Langem zu den Standardverfahren der Item Response Modelle. Der Test ist asymptotisch χ^2-verteilt mit $(k - 1)(g - 1)$ Freiheitsgraden (mit k Anzahl der Items und g Anzahl der Teilgruppen). Trotz seiner großen Bedeutung liegen bislang kaum Befunde hinsichtlich der Qualität der Approximation der Testgröße an die theoretische χ^2-Verteilung bei gegebenem Stichprobenumfang n und Itemzahl vor. In vorliegender Studie wird daher zuerst die Güte der Anpassung für endliches n und k untersucht. In Fällen, wo die Anpassung als nicht hinreichend zu beurteilen ist, steht der Bootstrap (Davison & Hinkley, 1997) als Methode der Wahl zur Gewinnung einer Prüfverteilung zur Verfügung. Dabei lassen sich grob der "naive" und der parametrische Bootstrap unterscheiden. Bei ersterem wird wiederholt aus den vorliegenden Beobachtungen mit Zurücklegen gezogen, während letzterer eine Monte-Carlo-Methode unter Verwendung der Modellgleichung und der geschätzten Item- und Personenparameter darstellt.
Im konkreten Fall des LRT stellt sich jedoch eine weitere Frage: Da mittels der aus theoretischen Gründen zu bevorzugenden Conditional Maximum Likelihood Methode (CML) primär nur Itemparameter gewonnen werden, ist nicht eindeutig geklärt, wie für den parametrischen Bootstrap die Personenparameter gewählt werden. Erstens kann – auf die Modelleigenschaft der spezifischen Objektivität zurückgreifend – aus einer beliebig (allerdings zumeist normal-)verteilten Zufallsverteilung gezogen werden, zweitens können die mittels Maximum Likelihood gewonnene Personenparameter verwendet werden und drittens kann auch ein sequentielles Wahrscheinlichkeitsverfahren gewählt werden, welches konstante Zeilenrandsummen gewährleistet.
Der Vergleich der drei Varianten zeigt, dass die drei Verfahren zu teilweise enorm abweichenden Ergebnissen führen können. Damit belegt diese Studie klar, daß die unbedachte Wahl des Bootstrapverfahrens grob irreführende Ergebnisse liefern kann. Ferner wird auch die Unanwendbarkeit des naiven Bootstraps schlüssig aufgezeigt.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38, 123–140.
Davison, A. C. and Hinkley, D. V. (1997). Bootstrap Methods and their Application. Cambridge: University Press.

Nichtparametrische Tests für das Rasch Modell bei kleinem Stichprobenumfang: Eine empirische Poweranalyse

Ingrid Koller[1]
Reinhold Hatzinger[2]
[1]Universität Wien
Institut für Psychologische Grundlagenforschung
Liebiggasse 5
1010 Wien

ingrid.koller@univie.ac.at

[2] Wirtschaftsuniversität Wien, Institut für Statistik und Mathematik
Augasse 2-6
1090 Wien

reinhold.hatzinger@wu-wien.ac.at

In der Praxis wäre es von Vorteil, Items eines Tests bereits anhand kleiner Stichproben hinsichtlich Rasch-Modell-Gültigkeit überprüfen zu können. Jedoch basieren viele Modelltests auf asymptotischen Eigenschaften, welche nur bei großen Stichproben zutreffen. Basierend auf der Idee des exakten Tests nach Rasch (1960) entwickelte Ponocny (1996, 2001) eine nichtparametrische Methode, die bereits bei kleinen Stichproben die Überprüfung auf Modellgültigkeit erlaubt. Die Ziele der vorliegenden Studie waren die Analyse der Typ-I Fehlerraten und die Bestimmung der empirischen Power für vier von Ponocny (2001) vorgeschlagenen Teststatistiken: T_4 und T_{10} für Differential Item Functioning (DIF), sowie T_2 und T_{11} zur Überprüfung der lokalen stochastischen Unabhängigkeit (lsU). Ebenso erfolgte eine Gegenüberstellung mit dem Likelihood Quotienten Test (LQT) nach Andersen (1973). Die Teststatistiken wurden anhand der modifizierten Markov-Chain-Monte-Carlo-Methode nach Verhelst (2008) in eRm (Mair & Hatzinger, 2007a; 2007b) berechnet. Jede Simulation war eine Funktion der Stichprobengröße (n = 30, n = 50, n =100, n = 200), der Testlänge (k = 5, k = 10, k = 20), der variierenden Anzahl an problematischen Items (k* = 1, k* = 2, k* = 5), sowie des Grades der Verletzung (r = 0.1 bis 0.9 bei lsU bzw. Itemparameterdifferenz = 0.1 bis 2 bei DIF, jeweils in 0.1 Schritten). Generell zeigen die Ergebnisse, dass die nonparametrischen Tests eine erste Modelleinschätzung bereits ab einen Stichprobenumfang von n = 50 erlauben, wobei eine optimale Power ab einem n = 200 erreicht werden kann. Daraus resultiert die Möglichkeit, bereits bei relativ kleinen Stichproben Rasch-Modell-Konformität bei Item-Sets zu überprüfen. Davon ausgehend wird eine neue Strategie zur Vorgehensweise bei Testkonstruktionen vorgeschlagen.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38(1), 123-140.

Mair, P., & Hatzinger, R. (2007a). CML based estimation of extended Rasch models with the eRm package in R. Psychology Science, 49(1), 26-43.

Mair, P., & Hatzinger, R. (2007b). Extended Rasch modeling. The eRm package for the application of IRT models in R. Journal of Statistical Software, 20(9). Retrieved from http://www.jstatsoft.org.

Ponocny, I. (1996). Kombinatorische Modelltests für das Rasch-Modell. [Combinatorial goodness-of-fit tests for the Rasch model.] (Unpublished doctoral thesis, University of Vienna, Austria).

Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. Psychometrika, 66(3), 437-460.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Kopenhagen: Danish Institute for Educational Research.

Verhelst, N. D. (2008). An efficient MCMC algorithm to sample binary matrices with fixed marginals. Psychometrika, 74(4), 705-728.

Hypothesentests bei Rasch Personen-Fit-Statistiken - Eine Alternative zum konventionellen Monte-Carlo-Verfahren

Christian Spoden
Jens Fleischer
Volker Zischka
Detlev Leutner

Universität Duisburg-Essen
Berliner Platz 6-8 (Weststadttürme)
45127 Essen

christian.spoden@uni-due.de

Bei Personen-Fit-Statistiken im Kontext der probabilistischen Testtheorie werden Standardisierungsformeln herangezogen, um mangelnde Modellpassung einzelner Antwort-Vektoren im Sinne von Hypothesentests mit festgelegtem α-Niveau zu untersuchen. Simulationsstudien belegen, dass die empirischen Verteilungen standardisierter Personen-Fit-Statistiken von der unterstellten Normalverteilung abweichen können, insbesondere wenn statt der wahren Item- und Personen-Parameter Schätzer in die Statistik eingehen (de la Torre & Deng, 2008; van Krimpen-Stoop & Meijer, 1999). Eine Alternative stellt die (parametrische) Monte-Carlo-Simulation der Verteilung dar, wie sie z.B. im R-Paket „ltm" implementiert ist und Datensätze unter der nicht ganz korrekten Annahme einer Normalverteilung der posterioren Fähigkeits-Verteilung simuliert (Rizopoulos, 2010). Für das dichotome Rasch-Modell könnte sich durch den Rasch Sampler (Verhelst, 2008) eine weitere Alternative ergeben. Dieser generiert binäre Daten-Matrizen, deren Randsummen mit der Ursprungsmatrix übereinstimmen. Unter Modelgültigkeit besitzen diese Matrizen gleiche Wahrscheinlichkeit, so dass nicht-parametrische Tests der beobachteten Datenmatrix möglich sind. Ebenso können aus den simulierten Datenmatrizen Antwort-Vektoren extrahiert werden, deren Rohsummenwerte jenen der beobachteten Matrix entsprechen. Es ergibt sich so die Möglichkeit zur Simulation von p-Werten bei Hypothesentests zur Prüfung der Modellpassung jedes einzelnen Antwort-Vektors im Sinne des Personen-Fits. Typ-I-Fehler und Power typischer Personen-Fit-Statistiken für das Rasch-Modell wurden unter diesem neuen Ansatz in zwei Simulationsstudien mit standardisierten Personen-Fit-Statistiken und einem konventionellen Monte-Carlo-Ansatz verglichen. Die Ergebnisse zeigen, dass (1) der empirische dem nominellen Typ-I-Fehler beim neuen Ansatz am besten entspricht, (2) die Power des neuen Ansatz in etwa der Power der standardisierten Statistiken gleicht und (3) der Vergleich verschiedener Personen-Fit-Statistiken innerhalb des neuen Ansatz Unterschiede zwischen Statistiken gemäß früheren Befunden bestätigt (vgl. Karabatsos, 2003).

de la Torre, J., & Deng, W. (2008). Improving person fit assessment by correcting the ability estimate and its reference distribution. Journal of Educational Measurement, 45, 159-177.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. Applied Measurement in Education, 16, 277-298.

Rizopoulos, D. (2010). ltm - Latent Trait Models under IRT. Reference manual.- URL: http://cran.r-project.org/web/packages/ltm/ltm.pdf (Download: 02.02.2011).

van Krimpen-Stoop, E. M. L. A. & Meijer, R. R. (1999).The null distribution of person-fit statistics for conventional and adaptive tests. Applied Psychological Measurement, 23, 164-180.

Verhelst, N. D. (2008) An efficient MCMC algorithm to sample binary matrices with fixed marginals. Psychometrika, 73, 705–728.

Do., 22.09., Anwendungsorientierte methodische Forschung II, 15.50-16.10 Uhr, Kapelle M3/126N (K)

Validation of a multidimensional willingness to communicate scale

Purya Baghaei

Islamic Azad University
Mashhad Branch
Iran

Willingness to communicate (WTC) is one of the affective factors assumed to influence success in second and foreign language learning. The purpose of the present study is to develop and validate a multidimensional WTC scale that can be used in foreign language learning contexts. A 22-item three-dimensional WTC scale was developed. The three dimensions included WTC with native speakers, WTC with non-native speakers and WTC in the classroom. The questionnaire was given to a sample of 287 Iranian students of English as a foreign language. Multidimensional and unidimensional Rasch models implemented in the ConQuest programme were used to compare the fits of three models: a unidimensional model, a two-dimensional model and a three-dimensional model. Results showed that the three-dimensional model significantly fits better than the other two models. The low and moderate correlations among the dimensions further supported the distinctness of the dimensions.

Do., 22.09., Anwendungsorientierte methodische Forschung II, 16.10-16.30 Uhr, Kapelle M3/126N (K)

Psychometrische Untersuchung der Hopkins Symptom Checklist 25 (HSCL-25) mit dem Partial Credit Modell

Herbert Poinstingl[1,2]
Philipp Yorck Herzberg[2,3]
Elmar Brähler[3]
Klaus D. Kubinger[2]

[1]Universität der Bundeswehr Hamburg
Fakultät für Geistes- und Sozialwissenschaften
Holstenhofweg 85
22043 Hamburg

herbert.poinstingl@hsu-hh.de

[2]Universität Wien, Fakultät für Psychologie

[3]Universität Leipzig, Medizinische Fakultät

Die Hopkins Symptom Checklist (HSCL) ist ein, insbesondere in der klinischen Psychologie und Psychiatrie, weitverbreitetes Verfahren, dass in verschiedenen Versionen mit unterschiedlicher Länge zur Erfassung von psychischen Symptomen verwendet wird. Die HSCL-25 ist eine Kurzversion der HSCL mit 25 Items, das als Screening-Instrument zur Erfassung von Angst und Depression eingesetzt wird. Im Rahmen einer von der USUMA im Auftrag der Universität Leipzig durchgeführten bevölkerungsrepräsentativen Befragung wurden in zwei Wellen N=2520 Personen befragt, wobei auch die HSCL-25 eingesetzt wurde.
In diesem Beitrag sollen die Ergebnisse einer Analyse mit dem Partial Credit Modell (PCM; Masters, 1982) und der anschließenden Überprüfung der PCM-Gültigkeit der HSCL-25 diskutiert werden. Zur Überprüfung der PCM-Gültigkeit kamen die folgenden Verfahren zur Anwendung: Likelihood Ratio Test (LRT; Andersen, 1973), Grafischer Modelltest (GMT; Rasch, 1960) und Paarvergleichstest nach Wald (Glas & Verhelst, 1995). Die testtheoretischen Analysen wurden mit dem R-Package „extended Rasch modeling" (eRm; Mair & Hatzinger, 2010) durchgeführt.

Do., 22.09., Anwendungsorientierte methodische Forschung II, 16.30-16.50 Uhr, Kapelle M3/126N (K)

Test- und Raterdesigns für das Assessment und die Kodierung freier Schreibaufgaben

Katrin Böhme
Stefan Schipolowski

Institut zur Qualitätsentwicklung im Bildungswesen (IQB)
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin

katrin.boehme@iqb.hu-berlin.de

Im Rahmen von Large-Scale Assessments werden nur selten Aufgaben eingesetzt, die von den Testteilnehmern umfangreiche freie Textproduktionen erfordern. Dies begründet sich unter anderem in der äußerst aufwändigen Auswertung freier Schreibprodukte durch geschulte Rater. Neben der Frage, welche Kriterien für die Beurteilung der Texte gewählt werden können, ist auch die Frage relevant, wie Ratereffekte, bspw. Strenge- bzw. Mildeeffekte oder Halo-Effekte, identifiziert und gehandhabt werden können. Für die Überprüfung der Interraterreliabilität sowie die Identifikation von Ratereffekten ist ausschlaggebend, in welcher Weise und in welchem Umfang Mehrfachkodierungen im Raterdesign vorgesehen werden.

Vor diesem Hintergrund wird im vorliegenden Beitrag ein Überblick über verschiedene Varianten von Test- und Raterdesigns für das Assessment und die Kodierung freier Schreibaufgaben gegeben. So ist bspw. für das Testdesign relevant, welche dimensionale Struktur für das Konstrukt der Schreibkompetenz angenommen wird und wie diese modelliert werden soll. Als Raterdesigns sind neben so genannten complete und incomplete block designs solche Designs denkbar, bei denen Rater innerhalb von Schülern oder Schüler innerhalb von Ratern genestet sind (vgl. Hoyt, 2000).

Die Art der gewählten Designs determiniert ferner die mit den Daten möglichen Analysen. Daher soll in einem zweiten Schritt auch auf einige klassische datenanalytische Zugänge für die Schätzung der Schreibkompetenz und die hierfür jeweils optimalen Designvarianten eingegangen werden. Relevant sind in diesem Zusammenhang traditionell bspw. Varianzkomponentenmodelle bzw. Modelle der Generalisierbarkeitstheorie (Brennan, 2001), Multifacettenmodelle (Weigle, 1998) oder Strukturgleichungsmodelle (Schoonen, 2005).

Abschließend werden die vorgestellten Möglichkeiten der Designgestaltung anhand realer Daten aus den Studien der Pilotierung von VERA-8 2011 sowie der Normierung der Bildungsstandards für den Kompetenzbereich Schreiben 2011 illustriert und die hier gewählten datenanalytischen Zugänge erläutert.

Brennan, R. L. (2001). Generalizability theory. New York: Springer. Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? Psychological Methods, 5, 64–86.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modelling. Language Testing, 22, 1–30.

Weigle, S. C. (1998). Using FACETS to model rater training effects. Language Testing, 15, 263–287.

TAGUNGSBEITRÄGE

Freitag

Fr., 23.09., Symposium: Kausale Effekte: Fortschritte in Theorie und Anwendung, 9.00-11.00 Uhr, Hörsaal M3/232N (H)

Symposium: Kausale Effekte: Fortschritte in Theorie und Anwendung

Benjamin Nagengast[1,2]

[1]University of Oxford
Department of Education
15 Norham Gardens
Oxford OX2 7PX
United Kingdom

[2]Universität Tübingen
Institut für Erziehungswissenschaft
Europastr. 6
72072 Tübingen

benjamin.nagengast
@education.ox.ac.uk

Die Analyse kausaler Effekte in Quasi-Experimenten findet zunehmend Anwendung in der Psychologie und anderen Sozialwissenschaften. Ausgehend von Theorien kausaler Effekte (z.B. Steyer et al., in press) werden dabei in der Regel durchschnittliche Behandlungseffekte dichotomer Treatments untersucht. In den letzten Jahren wurde jedoch der theoretische Rahmen auf Mediationseffekte, Mehrebenendesigns und komplexere Behandlungsvariablen erweitert. Auch Tests der Voraussetzungen für kausale Schlüsse und alternative statistische Analysemethoden finden zunehmend Interesse. Das Symposium gibt einen Überblick über diese aktuellen Entwicklungen: In den theoretischen Beiträgen von Axel Mayer und Felix Thoemmes, werden Effektdefinitionen für kausale Mediatoranalysen eingeführt und Kausalitätsbedingungen verschiedener Rahmenmodelle miteinander verglichen. Sonja Hahns Beitrag befasst sich mit Permutationstests bei vorhandenen Interaktionen, die für die Analyse kausaler Effekte in kleinen Stichproben relevant sind. Die Beiträge von Benjamin Nagengast und Ulf Kröhne haben einen angewandten Fokus und beschäftigen sich mit der Analyse von kausalen Effekten geordnet kategorialer Treatments und der Analyse von kausalen Effekten in Mehrebenendesigns in Large-Scale Assessment Studien. Die Beiträge werden abschließend von Rolf Steyer diskutiert.

Beiträge

1. Axel Mayer: Behandlungseffekte für Behandelte (Effects on the treated) in kausalen Mediationsmodellen.
2. Felix Thoemmes: Vergleich von Kausalitätsbedingungen.
3. Sonja Hahn: Vergleich von Permutations- und parametrischen Tests bei Treatment-Kovariaten-Interaktion.
4. Benjamin Nagengast: Adjustierungsverfahren für geordnet kategoriale Treatments: Ein theoretischer Vergleich und eine empirische Anwendung.
5. Ulf Kröhne: Beurteilung der Wirksamkeit bilingualen Unterrichts: Vergleich von Strategien zur Identifikation und von Verfahren zur Schätzung kausaler Effekte in Large-Scale Assessments am Beispiel einer Re-Analyse einer Teilstichprobe der DESI-Studie.
6. Rolf Steyer: Diskussion

Fr., 23.09., Symposium: Kausale Effekte: Fortschritte in Theorie und Anwendung, 9.00-9.20 Uhr, Hörsaal M3/232N (H)

Behandlungseffekte für Behandelte (Effects on the treated) in kausalen Mediationsmodellen

Axel Mayer
Helmut Kirchmann
Rolf Steyer

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl für Methodenlehre und Evaluationsforschung
Am Steiger 3, Haus 1
07743 Jena

Axel.Mayer@uni-jena.de

helmut.kirchmann@med.uni-jena.de

rolf.steyer@uni-jena.de

In Beobachtungsstudien erfolgt die Zuweisung von Personen zu einer Behandlungsbedingung in der Regel nicht zufällig, sondern kann zum Beispiel von Selbstselektionsmechanismen oder von Experteneinschätzungen abhängen. In diesem Fall können (bedingte) Behandlungseffekte für die Behandelten (sogenannte Effects on the treated) wichtige Aufschlüsse über die Wirksamkeit der Behandlung geben. Besonders interessant sind diese Effects on the treated im Kontext von Mediationsanalysen, da hier der durchschnittliche direkte Effekt von der Verteilung der Behandlungsvariablen abhängt. (Bedingte) Behandlungseffekte für Behandelte unterliegen dahingegen nicht dieser Einschränkung.

Aufbauend auf einer probabilistischen Kausalitätstheorie (Steyer et al. 2012, Mayer et al. 2011), diskutieren wir die Behandlungseffekte für Behandelte in komplexen Mediationsmodellen. Zur Analyse von totalen, direkten und indirekten Behandlungseffekten für Behandelte verwenden wir ein Mehrgruppen-Strukturgleichungsmodell. Das Vorgehen wird anhand eines Datensatzes aus der Psychotherapieforschung illustriert (Kirchmann et al. 2011). Dieses Beispiel beinhaltet neben latenten und manifesten Kovariaten, eine latente Mediator-Variable und eine latente Outcome-Variable. Interaktionen zwischen der Behandlungsvariablen und den Kovariaten beziehungsweise dem Mediator werden berücksichtigt.

Kirchmann, H., Steyer, R., Joraschky, P., Schreiber-Willnow, K., Strauss, B. (2011). Are long-term treatment effects of depression following impatient group psychotherapy mediated by self-report attachment characteristics? Manuscript under review.

Mayer, A., Thoemmes, F., Rose, N., Steyer, R., & West, S.G. (2011). Theory and analysis of total, direct and indirect causal effects. Manuscript under review.

Steyer, R., Partchev, I., Kröhne, U., Nagengast, B., & Fiege, C. (in press). Probability and causality. Heidelberg, Germany: Springer.

Vergleich von Kausalitätsbedingungen

Felix Thoemmes[1]
Axel Mayer[2]
Rolf Steyer[2]

[1]Universität Tübingen
Institut für Erziehungswissenschaft
Europastr. 6
72072 Tübingen

felix.thoemmes@gmail.com

[2]Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl für Methodenlehre und Evaluationsforschung
Am Steiger 3, Haus 1
07743 Jena

axel.mayer@uni-jena.de

rolf.steyer@uni-jena.de

In der Literatur zur Theorie kausaler Effekte wurden – teils stark voneinander isoliert – bestimmte Bedingungen postuliert, unter denen ein kausaler Effekt unverzerrt geschätzt werden kann. Einen Vergleich verschiedener Kausalitätstheorien und deren Annahmen (und letztendlich auch der daraus resultierenden Verfahren zur Adjustierung von Effekten) findet man in der Literatur kaum. In diesem Beitrag sollen zumindest drei verschiedene Ansätze verglichen werden. Zum einen die Theorie kausaler Effekte mit „potential outcomes“ (Rubin, 2005), die stochastische Theorie kausaler Effekte (Steyer, Partchev, Kröhne, Nagengast, & Fiege, in press), und die DAG (directed acyclic graphs) Theorie (Pearl, 2000). Es werden die folgenden Bedingungen und deren Implikation zur Selektion von Kovariaten betrachtet: die „ignorability assumption“ von Rubin, die „unconfoundedness assumption“ von Steyer, und das „back-door criterion“ von Pearl. Anhand von Beispielen soll untersucht werden, zu welchem Schluss bezüglich kausaler Effekte die drei verschiedenen Bedingungen kommen. Hierbei zeigt sich, dass „ignorability“ eine empirisch nicht testbare Behauptung ist, die weder verifiziert noch falsifiziert werden kann. Dies führt zur Empfehlung, immer bezüglich aller beobachteter Kovariaten zu adjustieren (Rubin, 2009). Die „unconfoundedness“ kann mit Hilfe gängiger Verfahren empirisch getestet (falsifiziert) werden. Das „back-door criterion“ kann ebenfalls empirisch getestet werden, setzt aber auch voraus dass sämtliche Kausalbeziehungen zwischen den Kovariaten expliziert werden können. An Beispielen soll veranschaulicht werden wie die verschiedenen Annahmen in einem angewandten Kontext benutzt werden können und zu welchen Schlüssen sie gelangen.

Pearl, J. (2000). Causality: models, reasoning, and inference. Cambridge University Press. Rubin, D. B. (2005). Causal inference using potential outcomes. Journal of the American Statistical Association, 100(469), 322-331.

Rubin, D. B. (2009). Author’s reply: Should observational studies be designed to allow a lack of balance in covariate distributions across treatment groups? Statistics in Medicine, 28(9), 1420-1423.

Steyer, R., Partchev, I., Kröhne, U., Nagengast, B., & Fiege, C. (in press). Probability and Causality 1. Springer.

Fr., 23.09., Symposium: Kausale Effekte: Fortschritte in Theorie und Anwendung, 9.40-10.00 Uhr, Hörsaal M3/232N (H)

Vergleich von Permutations- und parametrischen Tests bei Treatment-Kovariaten-Interaktion

Sonja Hahn

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl für Methodenlehre und Evaluationsforschung
Am Steiger 3, Haus 1
07743 Jena

hahn.sonja@uni-jena.de

Verfahren wie die ANCOVA werden u.a. verwendet, um eine statistische Kontrolle von Kovariaten zu ermöglichen. Dies ist beispielsweise notwendig, wenn ein Quasi-Experiment vorliegt, aber Hypothesen über den durchschnittlichen Effekt eines Treatments getestet werden sollen.
Sowohl bei gängigen parametrischen Tests (z.B. ANCOVA) als auch bei Permutationstests (vgl. Good, 2002, Kap. 4) gibt es verschiedene Ansätze für die statistische Kontrolle von Kovariaten. Viele Tests aus beiden Verfahrensgruppen setzen voraus, dass keine Interaktionen von Kovariaten und Treatment in Bezug auf die abhängige Variable existieren. Im Gegensatz dazu finden sich in einigen Anwendungsgebieten solche Interaktionen, z.B. Untersuchungen zum Expertise-Reversal-Effekt, in denen die Wirkung einer didaktischen Intervention vom Vorwissen abhängt.
Eine Untersuchung aus diesem Bereich (Pollock, Chandler und Sweller, 2002) dient als Grundlage für die vorliegende Simulationsstudie. In den simulierten Datensätzen ist zwar eine Interaktion, aber kein durchschnittlicher Effekt des Treatments vorhanden. Es wird untersucht, ob verschiedene Verfahren das Signifikanzniveau beim Testen des durchschnittlichen Effekts einhalten. Ein Schwerpunkt stellen Verfahren dar, die auf dem Prinzip der Permutation beruhen. Aufgrund zunehmender Rechnerleistungen gibt es ein wachsendes Interesse an dieser Verfahrensgruppe.

Good, P. (2000). Permutation Tests – A Practical Guide to Resampling Methods for Testing Hypotheses (2nd edition). New York: Springer.

Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. Learning and Instruction, 12, 61–86.

Adjustierungsverfahren für geordnet kategoriale Treatments: Ein theoretischer Vergleich und eine empirische Anwendung

Benjamin Nagengast[1,2]
Herbert W. Marsh[1]
Carlo Chiorri[1,3]
Kit-Tai Hau[4]

[1]University of Oxford
Department of Education
15 Norham Gardens
Oxford OX2 7PX
United Kingdom

benjamin.nagengast
@education.ox.ac.uk

herb.marsh
@education.ox.ac.uk

[z]Universität Tübingen
Institut für Erziehungswissenschaft
Europastr. 6
72072 Tübingen

[3] Universita di Genova
Dipartimento di Scienze Antropologiche
Sezione di Psicologia
Corso A. Podestà (IV piano)
16128 Genova, Italia

carlo.chiorri@gmail.com

[4] The Chinese University of Hongkong
Faculty of Education
Department of Educational Psychology Shatin, N.T., Hong Kong.

kthau@cuhk.edu.hk

Theorie und Verfahren zur Analyse kausaler Effekte sind in der Regel auf dichotome Treatments (z.B. den Vergleich einer Experimental- und einer Kontrollgruppe) beschränkt. In Anwendungen sind jedoch häufig auch die Effekte geordnet kategorialer oder kontinuierlicher Treatments von Interesse. Dazu existieren verschiedene Ansätze zur Adjustierung von Dose-Response-Kurven mit generalisierten Balancing-Scores, die inzwischen in leicht zugänglicher Software implementiert sind. Allerdings kommen diese in den Sozialwissenschaften nur selten zur Anwendung. Stattdessen wird häufig auf die unangemessene Dichotomisierung der Behandlungsvariablen zurückgegriffen.
Der Beitrag vergleicht vier Adjustierungsverfahren für geordnet kategoriale Treatments (Kovarianzanalyse, Subklassifizierung, Matching und Gewichtung) hinsichtlich ihrer theoretischen Grundannahmen und Adjustierungsprinzipien. Anhand eines Beispiels aus der empirischen Bildungsforschung (Schätzung des Effekts von Erwerbstätigkeit neben der Schule) werden die Effektschätzungen der verschiedenen Adjustierungsverfahren miteinander und mit herkömmlichen – auf Dichotomisierung beruhenden und daher weniger angemessenen – Verfahren verglichen.

Beurteilung der Wirksamkeit bilingualen Unterrichts: Vergleich von Strategien zur Identifikation und von Verfahren zur Schätzung kausaler Effekte in Large-Scale Assessments am Beispiel einer Re-Analyse einer Teilstichprobe der DESI-Studie

Ulf Kröhne
Johannes Hartig
Eckhard Klieme

Deutsches Institut für Internationale Pädagogische Forschung
Schloßstraße 29
60486 Frankfurt am Main

kroehne@dipf.de

hartig@dipf.de

klieme@dipf.de

Für die Studie Deutsch-Englisch-Schülerleistungen-International (DESI) soll in einer Re-Analyse die Wirksamkeit bilingualer Angebote zur Förderung sprachlicher Kompetenz beurteilt werden. Da die Teilnahme an bilingualen Angeboten im Rahmen der Large-Scale Studie nicht kontrolliert sondern nur beobachtet wurde, kann aus dem nicht-adjustierten Vergleich von Schülern mit und ohne bilingualem Unterricht nicht ursächlich auf die Wirksamkeit bilingualen Unterrichts geschlossen werden. In dem Beitrag wird deshalb ausgehend von den theoretischen Bedingungen zur Identifikation kausaler Effekte diskutiert, welche Annahmen zur Beurteilung der Wirksamkeit bilingualen Unterrichts benötigt werden. Dabei wird beispielhaft für die typische Datenlage in Large-Scale Assessments dargestellt, dass auf Schülerebene eine Unverfälschtheitsannahme nicht beibehalten werden kann, wenn theoretisch bedeutsame, zeitlich vorgeordnete Kovariaten nicht verfügbar sind. Deshalb wird für eine Teilstichprobe der DESI-Studie ein Falsifikationsversuch der Wirksamkeit bilingualen Unterrichts auf Schulebene vorgeschlagen. Der daraus abgeleitete und mit Hilfe einer Unverfälschtheitsannahme auf der Analyseebene Schule identifizierte durchschnittliche kausale Effekt kann wiederum mit Hilfe verschiedener Analyseverfahren geschätzt werden. Jedes Analyseverfahren benötigt bei der vorliegenden Datenlage zusätzliche Annahmen zur Adjustierung vorhandener Schulunterschiede. In dem Vortrag werden deshalb verschiedene Propensity-Score-basierte Adjustierungstechniken und Verfahren basierend auf der generalisierten Kovarianzanalyse zur Beurteilung der Wirksamkeit bilingualen Unterrichts auf Schulebene angewendet. Die geschätzten Effekte werden zwischen den Adjustierungsverfahren verglichen. Es zeigen sich übereinstimmend höhere sprachliche Kompetenzen bei Schülerinnen und Schülern, die an Schulen mit bilingualem Unterricht unterrichtet werden. Es wird geschlussfolgert, dass diese Unterschiede auf das Vorhandensein bilingualer Angebote an der Schule zurückgeführt werden können. In der abschließenden Diskussion werden die inhaltlichen und methodischen Einschränkungen der durchgeführten Re-Analyse kritisch reflektiert und es werden mögliche Generalisierungen des Vorgehens dargestellt.

Fr., 23.09., Symposium: Kausale Effekte: Fortschritte in Theorie und Anwendung, 10.20-10.40 Uhr, Hörsaal M3/232N (H)

Diskussion

Rolf Steyer

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl für, Methodenlehre und Evaluationsforschung
Am Steiger 3, Haus 1
07743 Jena

rolf.steyer@uni-jena.de

Fr., 23.09., Symposium: Psychometrische Probleme bei Antwort- und Itemformaten, 9.00-10.40 Uhr, Kapelle M3/126N (K)

Symposium: Psychometrische Probleme bei Antwort- und Itemformaten

Moritz Heene

Institut für Psychologie
Karl Franzens Universität
Graz, Maiffredygasse 12b
A-8010 Graz, Österreich.

moritz.heene@uni-graz.at

Psychometrische Probleme von Antwort- und Itemformaten stellen ein Hauptproblem für die valide Interpretation von Fragebogenstudien dar. Wie Forschungsergebnisse zeigen, können konstruktirrelevante Einflüsse von Antwort- und Itemformaten die Ergebnisse verzerren. Das Symposium hat zum Ziel, die psychometrischen Auswirkungen von verschiedenen Antwort- und Itemformaten aufzuzeigen und Lösungsvorschläge abzuleiten. Dabei wird auf psychometrische Probleme durch die Verwendung von sogenannten „vague quantifiers“ zur Erfassung von Verhaltenshäufigkeiten, durch die Verwendung von negativ gepolten Items zur Vermeidung von Aquieszenz, durch die Verwendung von bestimmten Antwortformaten sowie durch das Auftreten von response sets eingegangen. Die Ergebnisse zeigen im Wesentlichen, dass Veränderungen von Item- und Antwortformaten massive Auswirkungen auf die psychometrische Struktur von Fragebogen haben und theoretisch inhaltsgleiche Skalen psychometrisch nicht mehr vergleichbar sind.

Beiträge:

1. Moritz Heene: Itempolung und seine Effekte auf die Qualität von Ratingskalen: Eine Überprüfung mithilfe des Generalized Partial Credit Modells

2. Alexandra Zech: Auswirkungen von vague quantifiers auf die faktorielle Struktur von Fragebogen

3. Sven Hilbert: Dichotom, ordinal oder analog? Wie unterschiedliche Antwortformate psychometrische Eigenschaften von Fragebogenitems bestimmen.

4. Matthias Ziegler: Response Set oder Faking-Stil? Eine Analyse von Bewerberdaten

Fr., 23.09., Symposium: Psychometrische Probleme bei Antwort- und Itemformaten, 9.00-9.25 Uhr, Kapelle M3/126N (K)

Itempolung und seine Effekte auf die Qualität von Ratingskalen: Eine Überprüfung mithilfe des Generalized Partial Credit Modells.

Moritz Heene
Alexandra Zech
Sven Hilbert
Markus Bühner

[1]Institut für Psychologie
Karl Franzens Universität Graz
Maiffredygasse 12b
A-8010 Graz, Österreich

moritz.heene@uni-graz.at

Zur Vermeidung von Aquieszenz enthalten viele Persönlichkeitsfragebogen positiv und negativ gepolte Items, wobei angenommen wird, dass beide Itemtypen äquivalente Indikatoren derselben latenten Variable sind. In faktorenanalytischen Untersuchungen zeigen sich allerdings durch die Itempolung konsistent zwei Faktoren, die durch die Aufteilung in positiv gegenüber negativ gepolten Items gekennzeichnet sind. Bisherige Forschung dazu liefert jedoch keine direkten Informationen über den Einfluss der Itempolung auf die psychometrische Qualität der Ratingskalen wie es Item-Response-Modelle ermöglichen. In einer Fragebogenstudie wurden daher die Daten von N = 1237 Probanden mit dem Generalized-Partial-Credit-Modell analysiert. Die Ergebnisse zeigen, dass es durch die Verwendung von negativ im Vergleich zu positiv gepolten Items zu starken Verzerrungseffekten hinsichtlich der Ordnung und Relation der Antwortkategorien kommt und die Äquivalenz beider Itemtypen nicht gegeben ist. Implikationen für die Testkonstruktion werden diskutiert.

Fr., 23.09., Symposium: Psychometrische Probleme bei Antwort- und Itemformaten, 9.25-9.50 Uhr, Kapelle M3/126N (K)

Auswirkungen von vague quantifiers auf die faktorielle Struktur von Fragebogen

Alexandra Zech[1]
Moritz Heene[1]
Sven Hilbert[1]
Stella Bollmann[2]
Markus Bühner[1]

[1]Institut für Psychologie
Karl Franzens Universität Graz
Maiffredygasse 12b
A-8010 Graz, Österreich

[2]Ludwig-Maximilians-Universität München
Leopoldstr. 13a
80802 München

alexandra.zech@uni-graz.at

moritz.heene@uni-graz.at

sven.hilbert@uni-graz.at

stella.bollmann@campus.lmu.de

markus.buehner@uni-graz.at

In Persönlichkeitsfragebogen werden oft ungefähre Häufigkeitsangaben von Verhaltensweisen in Form von sogenannten „vague quantifiers“ (z.B. „häufig“, „viel“ oder „manchmal“) verwendet. Allerdings unterscheiden sich Personen hinsichtlich der Interpretation von vague quantifiers. Bisherige Forschung beschäftigte sich daher intensiv mit der Validität der Häufigkeitsangaben, der Interpretation der Antwortskala in Abhängigkeit von der Art der vague quantifiers sowie deren Auswirkung auf die Erfassung von Gruppenunterschieden. Bislang ungeklärt ist allerdings die Auswirkung von vague quantifiers auf die faktorielle Struktur von Fragebogen. In der vorliegenden Studie wurde dazu einer Stichprobe von insgesamt N = 881 Probanden zwei verschiedene Versionen desselben Fragebogens vorgegeben. In einer Version wurden vague quantifiers benutzt, in der anderen wurden dieselben Fragen ohne jene vorgegeben. Mit einer Multigruppenfaktorenanalyse wurde der Effekt auf die Äquivalenz von verschiedenen Faktorstrukturparametern untersucht. Die Hauptbefunde weisen darauf hin, dass hinsichtlich der Faktorstukturen deutliche Unterschiede zwischen den beiden Fragebogenversionen bestehen und die jeweiligen Faktoren nicht miteinander vergleichbar sind.

Fr., 23.09., Symposium: Psychometrische Probleme bei Antwort- und Itemformaten, 9.50-10.15 Uhr, Kapelle M3/126N (K)

Dichotom, ordinal oder analog? Wie unterschiedliche Antwortformate psychometrische Eigenschaften von Fragebogenitems bestimmen.

Sven Hilbert[1]
Moritz Heene[1]
Alexandra Zech[1]
Matthias Ziegler[2]
Markus Bühner[1]

[1]Institut für Psychologie
Karl Franzens Universität Graz
Maiffredygasse 12b
A-8010 Graz, Österreich

sven.hilbert@uni-graz.at

moritz.heene@uni-graz.at

alexandra.zech@uni-graz.at

markus.buehner@uni-graz.at

[2]Humboldt-Universität zu Berlin
Institut für Psychologie
Unter den Linden 6
10099, Berlin

zieglema@hu-berlin.de

Für psychometrische Testverfahren existiert eine große Anzahl verschiedener Antwortformate. Diese dienen unter anderem dazu, Beantwortenden die Möglichkeit zu geben, in ihren Angaben gut differenzieren zu können ohne von den Auswahlmöglichkeiten überlastet zu werden. In Studien wurde bereits mehrfach gezeigt, dass sich mit der Zahl der Antwortmöglichkeiten auch die Reliabilität erhöht. In der vorliegenden Studie wurde nun untersucht, in welcher Form sich das Antwortformat auf die Kriteriumsvalidität, faktorielle Struktur und weitere psychometrische Kennwerte auswirkt. Hierfür bearbeiteten N = 866 Versuchspersonen drei Fragebogen mit identischen Items bezüglich der Persönlichkeitsdimension Gewissenhaftigkeit, jeweils einmal mit dichotomem und fünfstufigem Antwortformat sowie einer visuellen Analogskala. Als Kriterium diente die Anzahl der unentschuldigten Fehltage im letzten Schuljahr vor Studienbeginn. Wie erwartet führten die Antwortformate zu Unterschieden hinsichtlich mehrerer psychometrischer Kennwerte. Die Faktoren aus den Items mit den drei unterschiedlichen Antwortformaten zeigten außerdem nicht hinreichend hohe Faktorkorrelationen, den Einfluss des Antwortformats abermals unterstreichend. Hinsichtlich der Kriteriumsvalidität ergaben sich allerdings keine Unterschiede zwischen den drei Antwortformaten. Mögliche Erklärungsansätze werden diskutiert.

Fr., 23.09., Symposium: Psychometrische Probleme bei Antwort- und Itemformaten, 10.15-10.40 Uhr, Kapelle M3/126N (K)

Response Set oder Faking-Stil? Eine Analyse von Bewerberdaten

Matthias Ziegler[1]
Erik Danay[1]
Amy Gammon[2]
Richard Griffith[2]

[1]Humboldt-Universität zu Berlin
Institut für Psychologie
Unter den Linden 6
10099 Berlin

zieglema@hu-berlin.de

erik.danay@hu-berlin.de

[2]Florida Tech
I/O Psychology
150 W. University Blvd.
Melbourne, Florida 32901

griffith@fit.edu

gammona@gmail.com

Die Analyse von Fragebogendaten mithilfe des Mixed-Rasch-Models (MRM) hat in zahlreichen Untersuchungen eine Zweiklassenlösung ergeben. Allerdings wurden die Klassen je nach Erhebungsbedingungen unterschiedlich interpretiert. Die Ergebnisse aus Daten, die unter anonymen Bedingungen erhoben wurden, werden oft als Response Sets im Sinne von Mittel- und Extremkreuzern interpretiert. Bei Daten aus Bewerbungskontexten spricht man jedoch häufig von unterschiedlichen Faking-Stilen, nämlich Slight und Extreme Faking. Ob es zwischen diesen beiden Klassenlösungen Beziehungen gibt, wurde bisher nicht untersucht. In der vorliegenden Studie wurden Daten aus realen Bewerberkontexten analysiert (N = 497). Das besondere an den Daten ist, dass für alle Bewerber auch Daten derselben Fragebogen aus einem anonymen Kontext vorliegen. Mithilfe von MRMs wurden die Daten zunächst getrennt nach Erhebungsbedingung analysiert. Es zeigten sich erwartungskonform die Response Sets Mittel- und Extremkreuzen in der anonymen Bedingung und die Faking-Stile Slight und Extreme Faking im Bewerbungskontext. Zwischen den beiden Klassenlösungen besteht ein erstaunlich hoher Zusammenhang. So werden Extremkreuzer als Extreme Faker klassifiziert und Mittelkreuzer als Slight Faker. Nur ein kleiner Prozentsatz an Personen wechselt zwischen den Kategorien. In weiteren Analysen werden diese 3 Gruppen in Bezug auf die Variablen Big 5, Locus of Control, Integrität und Kontraproduktives Verhalten verglichen. Ergebnisse und deren praktische sowie theoretische Konsequenzen werden diskutiert.

Fr., 23.09., Einzelbeitrag, 10.40-11.00 Uhr, Kapelle M3/126N (K)

A framework for test equating based on the concepts of equity and causal effects

Safir Yousfi

German Federal Employment Agency
Psychological Research and Development

A formal approach to test equating relying on the concept of equity is proposed. It is shown that other equating criteria discussed in the literature (equal reliability, equal construct, symmetry, population invariance) are either inappropriate or implications of equity. It is argued that Lord's equity theorem does not justify serious doubts on the usefulness of the equity concept, in spite of the fact that there is no way to circumvent the theorem by refined techniques like local equating. Equating could be paraphrased as a technique of generating transformed scores for which the choice of the test form has no (substantial) causal effect for each individual. The result of this transformation might depend not only on the raw score but also on covariates and random noise. The relevance of the counterfactual perspective of causality for analyzing equating designs and techniques is outlined, as well as benefits of equating concepts and techniques for the analysis of causal effects.

Fr., 23.09., Nichtlineare Strukturgleichungsmodellierung, 14.00-14.20 Uhr, Hörsaal M3/232N (H)

Analyse von Moderatoreffekten mit Strukturgleichungsmodellen oder Partial Least Squares: Konsequenzen des Orthogonalisierens von Produkttermen

Christina S. Werner[1]
Karin Schermelleh-Engel[2]

[1]Freie Universität Berlin,
Habelschwerdter Allee 45
14195 Berlin

c.s.werner@fu-berlin.de

c.s.werner@psych.uni-frankfurt.de

[2]Johann Wolfgang Goethe-Universität
Mertonstraße 17
60054 Frankfurt am Main

schermelleh-engel@psych.uni-frankfurt.de

Für die Analyse von Moderatoreffekten zwischen latenten Variablen sind Produktindikator-Modelle recht verbreitet und einfach umsetzbar: Hierbei werden Indikatorvariablen der latenten Prädiktoren miteinander multipliziert, um einen latenten Produktterm zu operationalisieren. Solche Modelle können sowohl mit der in Strukturgleichungsmodellen gängigen Maximum Likelihood-Schätzung (ML) analysiert werden, als auch mittels Partial Least Squares-Schätzung (PLS). Ein potentielles Problem dieser Modelle ist jedoch die hohe Kollinearität von Produkttermen mit den Ausgangsvariablen. Sowohl für ML-Schätzung, als auch für PLS wurden daher Ansätze vorgeschlagen, bei denen die Kollinearität durch Orthogonalisierung der Produktvariablen eliminiert wird, d.h. durch Transformationen, die die Produkte von den übrigen Indikatorvariablen der Prädiktoren unabhängig machen. Solche Transformationen können allerdings für die Schätzmethoden unterschiedliche Konsequenzen haben.

Im Rahmen einer Simulationsstudie wurden daher ML- und PLS-Ansätze mit und ohne Orthogonalierung der Produktindikatoren untersucht. Bei nicht normalverteilten Prädiktorvariablen zeigten PLS-Analysen ohne Orthogonalisierung verzerrte Parameterschätzungen und überhöhte Alpha-Fehlerraten für den Test des Moderatoreffekts. Orthogonalisieren der Produktindikatoren reduzierte die Probleme für PLS deutlich. ML-Analysen zeigten dagegen bei Verwendung der untransformierten Produktindikatoren keine Probleme, während erst durch die Orthogonalisierung Verzerrungen der Schätzungen und Tests auftraten.

ML-Schätzungen unter Nutzung des nicht-orthogonalisierenden Unconstrained-Ansatzes waren generell robust gegenüber den untersuchten nichtnormalen Verteilungen in Bezug auf Parameterschätzungen, Teststärke und Alpha-Fehlerraten. Obwohl PLS-Schätzungen im Gegensatz zu ML nicht auf Verteilungsvoraussetzungen beruhen, waren sie dennoch anfällig gegenüber nichtnormalen Verteilungsformen. Insgesamt kann bei der Analyse von Moderatoreffekten in Produktindikator-Modellen eine Orthogonalisierung für varianzbasierte Methoden wie PLS empfehlenswert sein, für kovarianzbasierte ML-Schätzungen dagegen problematisch.

Fr., 23.09., Nichtlineare Strukturgleichungsmodellierung, 14.20-14.40 Uhr, Hörsaal M3/232N (H)

Chi-Quadrat Differenztest in nichtlinearen Strukturgleichungsmodellen

Carla Gerhard[1]
Karin Schermelleh-Engel[1]
Christina S. Werner[2]
Helfried Moosbrugger[1]

[1]Johann Wolfgang Goethe-Universität
Mertonstraße 17
60054 Frankfurt am Main

Gerhard@psych.uni-frankfurt.de

[2]Freie Universität Berlin
Habelschwerdter Allee 45
14195 Berlin

Bisher ist es nicht möglich, die Gesamtgüte eines nichtlinearen Strukturgleichungs-modells (SEM) zu bestimmen, da das benötigte Vergleichsmodell (saturiertes Modell) noch nicht definiert wurde. Das wirft die Frage auf, ob überhaupt bestimmte Gütekriterien linearer SEM im Rahmen nichtlinearer SEM anwendbar sein könnten. Verschiedene Charakteristika nichtlinearer SEM erschweren allerdings die Anwendung der bestehenden Gütekriterien bei gegebener Nichtlinearität. Dazu gehört u.a. die nichtnormale Verteilung der nichtlinearen Terme, die dazu führt, dass die gemeinsame Verteilung der Indikatoren multivariat nichtnormal ist. Die Anwendungsvoraussetzung vieler bestehender Gütekriterien wird durch diese Verteilungsform verletzt.
Ob der Chi-Quadrat-Differenztest zur Überprüfung einer oder mehrerer nichtlinearer Effekte geeignet ist, oder ob die Satorra-Bentler-Korrektur hier angemessener wäre, wurde bisher noch nicht untersucht.
Es wurde eine Simulationsstudie durchgeführt, um zu prüfen, unter welchen Bedingungen der normaltheoretische Chi-Quadrat Differenztest sowie Satorra-Bentler korrigierte Versionen dieses Tests verwendet werden können, um einen oder mehrere nichtlineare Effekte auf Signifikanz zu prüfen. Die Ergebnisse der Power- und Alpha-Fehler-Analysen werden berichtet.

Fr., 23.09., Nichtlineare Strukturgleichungsmodellierung, 14.40-15.00 Uhr, Hörsaal M3/232N (H)

Zum Problem der korrekten Standardisierung von nichtlinearen Strukturgleichungsmodellen

Karin Schermelleh-Engel[1]
Christina S. Werner[2]
Jana C. Gäde[1]
Carla Gerhard[1]
Helfried Moosbrugger[1]

[1]Johann Wolfgang Goethe-Universität
Mertonstraße 17
60054 Frankfurt am Main

schermelleh-engel
@psych.uni-frankfurt.de

[2]Freie Universität Berlin
Habelschwerdter Allee 45
14195 Berlin

Programme zur Analyse von linearen Strukturgleichungsmodellen (SEM) stellen routinemäßig eine komplett standardisierte Lösung zur Verfügung. Für nichtlineare SEM mit Produktindikatoren ist diese jedoch in der Regel nicht korrekt, da die Produktvariablen unabhängig von den in sie eingehenden Faktoren auf eine Varianz von eins standardisiert werden. Wen, Marsh und Hau (2010) konnten anhand eines Produkt-Indikator-Ansatzes (Unconstrained Approach) zeigen, dass aber unter Verwendung der vorhandenen, nicht korrekt standardisierten Parameter für Moderatoreffekte die korrekt standardisierten Parameter berechnet werden können. Ob diese Methode ohne Probleme auch auf Modelle angewandt werden kann, die neben einem Moderatoreffekt auch noch quadratische Effekte enthalten, wurde bisher nicht untersucht. Ein weiteres Problem besteht darin, dass nicht klar ist, ob die Wahl der Skaliervariable für die latenten nichtlinearen Terme einen Einfluss auf die Parameterschätzungen und damit auf die standardisierten Werte hat. Wen, Marsh und Hau (2010) schlugen vor, jeweils den Indikator mit der höchsten Reliabilität als Skalierer zu verwenden.

Untersucht werden sollte daher im Rahmen einer Monte Carlo-Studie anhand eines nichtlinearen SEM mit Moderator- und quadratischen Termen der Einfluss unterschiedlich reliabler Skalierer auf die Güte der Schätzung der nichtlinearen Parameter sowie die latenten Varianzen und Kovarianzen sowie auf Güte der standardisierten Lösung.

In der Simulationsstudie wurden Daten für ein Modell mit einem Moderatoreffekt (.20) und zwei quadratischen Effekten (.15, .10) generiert und die Reliabilität der drei Indikatorvariablen pro latentem Prädiktor variiert (.80, .60, .30). Analysiert wurden die Daten mit zwei LISREL-Ansätzen, dem Extended Unconstrained Approach (EUA), in welchem alle Parameter frei geschätzt werden, und dem Extended Constrained Approach (ECA), in welchem die nichtlinearen Parameter restringiert werden. Zum Vergleich wurde mit Latent Moderated Structural Equations (LMS) ein Ansatz verwendet, der keine Produktindikatoren benötigt.

Unsere Ergebnisse zeigen, dass die Reliabilität der Skalierer einen erheblichen Einfluss auf die Güte der Parameterschätzungen der Produkt-Indikator-Ansätze hat, vor allem auf die nichtlinearen Effekte des EUA. Beim ECA wurden die Populationswerte zwar ebenfalls nicht korrekt geschätzt, die Standardisierung führte dann aber zu den korrekten Werten. Als einziger Ansatz weist LMS keinerlei Schätzprobleme auf. Mögliche Ursachen für diese Ergebnisse werden diskutiert.

Fr., 23.09., Nichtlineare Strukturgleichungsmodellierung, 15.00-15.20 Uhr, Hörsaal M3/232N (H)

Ein neues Verfahren zur Schätzung von latenten nichtlinearen Effekten bei nicht-normalverteilten latenten Prädiktoren

Augustin Kelava[1]
Benjamin Nagengast[2,3]

[1]Technische Universität Darmstadt
Institut für Psychologie
Alexanderstr. 10
64283 Darmstadt

tino@augustin-kelava.de

[2]University of Oxford
Department of Education
15 Norham Gardens
Oxford OX2 7PX
United Kingdom

benjamin.nagengast@education.ox.ac.uk

[3]Universität Tübingen
Institut für Erziehungswissenschaft
Europastr. 6
72072 Tübingen

In den vergangenen 15 Jahren wurden zahlreiche Verfahren zur Schätzung von latenten nichtlinearen Interaktions- und quadratischen Effekten vorgestellt. Darunter fallen Produkt-Indikator Ansätze (z.B. Marsh, Wen & Hau, 2006), sog. distribution-analytic Ansätze (Klein & Moosbrugger, 2000, Klein & Muthèn, 2007) u.v.m. Alle diese Ansätze gehen von normalverteilten latenten Prädiktoren aus. Kelava und Nagengast (in Revision) stellten einen Bayesschen Ansatz vor, der die latente nichtnormale Prädiktorverteilung approximiert und den obigen Anätzen hinsichtlich Unverzerrtheit, Teststärke und Typ I Fehlerrate überlegen ist. In diesem Vortrag wird der Transfer des Verfahrens auf Nicht-Bayessche Modelle vorgestellt. In einer kleinen Simulationsstudie wird er mit gängigen Verfahren verglichen und seine Überlegenheit unter variierenden Bedingungen der Nicht-Normalität präsentiert. Vorteile und Limitation des neunen Verfahrens werden diskutiert.

Kelava, A. & Nagengast, B. (in Revision). A Bayesian model for the estimation of latent interaction and quadratic effects when latent variables are non-normally distributed. Multivariate Behavioral Research.

Klein, A. G. & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. Psychometrika, 65, 457-474.

Klein, A. G. & Muthèn, B. (2007). Quasi maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. Multivariate Behavioral Research, 47, 647-674.

Marsh, H. W., Wen, Z. & Hau., K.-T. (2006). Structural equation models of latent interaction and quadratic effects. In G. R. Hancock und R. O. Mueller (Hrsg.), Structural equation modeling: A second course (S. 225-265). Greenwich: Information Age Publishing.

Untersuchungen zu Rasch-Homogenität und response sets im Beck Depressions-Inventar Revision (BDI-II)

Ferdinand Keller[1]
Rainer W. Alexandrowicz[2]

[1]Klinik für Kinder- und Jugendpsychiatrie/Psychotherapie
Universität Ulm
Steinhövelstr. 5
89075 Ulm

ferdinand.keller
@uniklinikum-ulm.de

[2]Universität Klagenfurt

Hintergrund: Das BDI-II ist ein Selbstbeurteilungsinstrument zur Erfassung der Schwere depressiver Symptomatik und wird national wie international in zahlreichen klinischen Studien zur Depression eingesetzt. Inzwischen liegen viele Studien zur faktoriellen Struktur des BDI-II vor, während psychometrische Analysen mit probabilischen Testmodellen selten geblieben sind. Angesichts zunehmender Verbreitung des BDI-II über den klinischen Bereich hinaus erscheint es zudem interessant zu analysieren, wie gut das BDI-II bei Gesunden misst. Auch die psychometrischen Eigenschaften bei nicht primär depressiven Patienten sind noch unzureichend untersucht.
Fragestellung: wie gut und wie vergleichbar misst das BDI-II Depressivität in verschiedenen Stichproben (spezifische Objektivität)? Gilt das Rasch-Modell für die gesamte Personenstichprobe, oder gilt es mit jeweils unterschiedlichen Modellparametern nur in verschiedenen, unbekannten Teilstichproben (mixed Rasch-Modell (MRM))?
Methodik: Zur Analyse wurden drei Stichproben verwendet. Neben den beiden Manualstichproben, bestehend aus 266 Depressiven und insgesamt 582 Personen aus Normalstichproben, standen noch die Fragebögen von 898 Patienten, die in einer psychosomatischen Klinik in Behandlung waren, zur Verfügung. Die Analysen zum mixed Rasch-Modell wurden mit dem Programm WINMIRA v1.45 (v.Davier, 2001) durchgeführt.
Ergebnis: Die Ergebnisse lassen sich zusammenfassend so interpretieren, dass die Modellannahmen des Rasch-Modells erfüllt zu sein scheinen für die Stichproben der Depressiven und der Gesunden. Die Gleichheit der Itemparameter in beiden Stichproben muss jedoch bei einigen Items in Frage gestellt werden. Hinweise auf Heterogenität ergeben sich für die Patienten aus einer psychosomatischen Klinik. Eine Inspektion der Schwellen- und Personenparameter der 2-Klassenlösung des MRM verweist auf „response sets“, nämlich eine Gruppe von Personen mit Tendenz zur Mitte und eine Gruppe von „Extremankreuzern“. Den Zusammenhängen mit potentiell konfundierten Variablen, v.a. Diagnose, aber auch Alter und Geschlecht, wird derzeit nachgegangen.

Kontrolle von Antwortstilen durch die Anwendung von Raschmodellen und ihre Auswirkung auf Skaleninterkorrelationen

Jörg Heine
Alexandra Langmeyer
Christian Tarnai
Florian Hartmann

Universität der Bundeswehr München
Werner-Heisenberg-Weg 39
85577 Neubiberg

joerg-henrik.heine@unibw.de

alexandra.langmeyer@unibw.de

christian.tarnai@unibw.de

Florian.Hartmann@unibw.de

Individuelle Antwortstile im Sinne einer Verzerrungstendenz zu mittleren (MRS) oder extremen (ERS) Antwortkategorien sind bei mehrstufigen Antwortskalen ein häufig beobachtetes Phänomen. Im Rahmen der Skalierung mit Rasch und Mixed-Rasch Modellen ergeben sich beim Vorliegen solcher Antwortstile dann oft Mehrklassenlösungen. Die Verwendung der auf dieser Grundlage berechneten Personenparameter sollte zu Interkorrelationen von (Sub-)Skalen führen, die von den möglichen Verzerrungen durch die Antwortstile frei sind, und daher denen nach einer Theorie erwarteten Korrelationen eher entsprechen.
In der vorliegenden Untersuchung wird die Auswirkung der Berücksichtigung von Antwortstilen auf die Verbesserung der strukturellen Validität des Hexagonalen Modells der Interessen von Holland (1997) untersucht. Die theoriekonformen Korrelationen der sechs Orientierungen (R-realistic, I-investigative, A-artistic, S-social, E-enterprising und C-conventional) können durch einen Circumplex repräsentiert werden (Nagy et al., 2009).
Datenbasis für die durchgeführten Analysen sind zwei Stichproben von Studierenden verschiedener Fächer (N = 734 und N = 1283), bei denen die beruflichen Interessen nach dem Modell von Holland mit dem Allgemeinen Interessen-Struktur-Test (Bergmann & Eder, 2005) erfasst wurden.
Es zeigt sich, dass bei einzelnen Skalen das Mixed-Rasch Modell mit zwei Klassen besser zu den Daten passt und sich diese Klassen im Sinne der Auswirkung eines extremen Antwortstils (ERS) interpretieren lassen. Die Berücksichtigung der Personenparameterschätzungen an Stelle der Summenscores führt zu einer Veränderung der Interkorrelationen in Richtung einer besseren Entsprechung mit der circumplexen Struktur.
Die Ergebnisse werden detailliert auch in Hinblick auf die beiden verwendeten Programme mixRasch und Winmira und deren Parameterschätzmethoden jML und cML diskutiert.

Bergmann, C. & Eder, F. (2005). AIST-R - Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R). Göttingen: Beltz.

Holland, J.L. (1997). Making vocational choices. A theory of vocational personalities and work environments. Lutz, FL: Psychological Assessment Resources.

Nagy, G., Marsh, H. W., Lüdtke, O., & Trautwein, U. (2009). Representing the circles in our minds: Confirmatory factor analysis of circumplex structures and profiles. In T. Teo & M. S. Khine, Structural Equation Modelling in Educational Research: Concepts and Applications (pp. 287-315). Rotterdam, Netherlands: Sense Publishers

Zusammenlegung von Antwortkategorien bei der Untersuchung von Antwortstilen

Eunike Wetzel
Claus H. Carstensen

Otto-Friedrich-Universität Bamberg
Wilhelmsplatz 3
96047 Bamberg

eunike.wetzel@uni-bamberg.de

claus.carstensen@uni-bamberg.de

Forschung zum Thema Antwortstile zeigt, dass nicht alle Probanden die Antwortskala in der gleichen Weise verwenden, sondern es z.B. Personen gibt, die extreme Antworten bevorzugen, während andere diese vermeiden (Rost, Carstensen & von Davier, 1999; Austin, Deary & Egan, 2006). In Studien, die Antwortstile v.a. im Bereich der Persönlichkeitserfassung mittels Fragebögen untersuchen, wird häufig die mittlere Antwortkategorie mit der darunterliegenden zusammengelegt. Dies erzeugt die oftmals gewünschten geordneten Schwellenparameter und erleichtert damit die Interpretation der Daten. Allerding könnte diese Praxis die Analyse des Antwortverhaltens verzerren, da Personen, die unterschiedliche Antwortkategorien wählen, wie eine Gruppe behandelt werden, die sich hinsichtlich ihrer Traitausprägung nicht unterscheidet. In diesem Beitrag soll anhand einer Re-Analyse des NEO-PI-R (Ostendorf & Angleitner, 2004) untersucht werden, ob die Praxis der Zusammenlegung zweier Antwortkategorien gerechtfertigt ist.

Zunächst wurden Mixed Rasch Modelle in Winmira berechnet, um latente Klassen zu identifizieren, die sich als Antwortstile interpretieren lassen. Auf den meisten Facetten traten zwei Antwortstile konsistent auf, nämlich Extremkreuzer und Mittelkreuzer. Dann wurden in ConQuest Partial Credit Modelle berechnet. Die resultierenden Weighted Likelihood Estimates (WLE; Warm, 1989) für die einzelnen Antwortkategorien wurden verglichen. Es zeigten sich deutliche Unterschiede zwischen den WLEs für die einzelnen Antwortkategorien, auch zwischen den WLEs für die zweite und dritte Antwortkategorie. Weiterhin waren die WLEs für die meisten Items über alle Kategorien geordnet, d.h. hier schien die mittlere Kategorie auch die mittlere Traitausprägung zu messen. Diese Ergebnisse legen nahe, dass zwischen Personen, die die zweite vs. die dritte Antwortkategorie wählen, deutliche Unterschiede in ihrer Traitauspägung bestehen, so dass ihre Zusammenlegung in eine gemeinsame Gruppe nicht gerechtfertigt scheint.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. Personality and Individual Differences, 40(6), 1235–1245.

Ostendorf, F., & Angleitner, A. (2004). NEO-PI-R: NEO-Persönlichkeitsinventar nach Costa und McCrae. Göttingen: Hogrefe.

Rost, J., Carstensen, C. H., & von Davier, M. (1999). Sind die Big Five Rasch-skalierbar? - Eine Reananlyse der NEO-FFI-Normierungsdaten [Are the Big Five Rasch scalable? A reanalysis of the NEO-FFI norm data]. Diagnostica, 45(3), 119–127.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 45(3), 427-450.

IRT-Analyse von Traitausprägung und Antwortstilen in Ratingdaten

Thorsten Meiser[1]
Ulf Böckenholt[2]

[1] Fachbereich Psychologie
Universität Mannheim
Schloss Ehrenhof Ost
D-68131 Mannheim

t.meiser
@uni-mannheim.de

[2] Kellogg School of Management
Northwestern University
2001 Sheridan Road
Evanston, IL 60208, USA

u-bockenholt@kellogg.
northwestern.edu

Üblicherweise wird die Wahl einer Antwortkategorie auf einer mehrstufigen Ratingskala als Ausdruck der Stärke eines zugrundeliegenden Persönlichkeitsmerkmals interpretiert, das mit Hilfe der vorgegebenen Items erfasst werden soll. Zur Messung des Persönlichkeitsmerkmals eignen sich nach dieser Vorstellung eindimensionale IRT-Modelle wie das *Partial Credit*-Modell, das *Rating Scale*-Modell oder das *Graded Response*-Modell. In diesen eindimensionalen Modellen bleibt jedoch die Rolle von Antwortstilen unberücksichtigt, die neben dem zu erfassenden Trait einen wesentlichen Einfluss auf die Auswahl von Antwortkategorien ausüben und die Messung des angezielten Persönlichkeitsmerkmals verzerren können (etwa De Jong, Steenkamp, Fox & Baumgartner, 2008; Weijters, Geuens & Schillewaert, 2010). Daher wird in diesem Beitrag eine mehrdimensionale Modellierung von Ratingantworten illustriert, die neben interindividuellen Differenzen auf der inhaltlichen Traitdimension auch interindividuelle Differenzen in der Nutzung des Antwortformats beinhaltet (Böckenholt, under review). Die Spezifikation multidimensionaler IRT-Modelle erfolgt dabei als *Nonlinear Mixed Models* (Rijmen, Tuerlinckx, de Boeck & Kuppens, 2003), die die zusätzliche Analyse des Einflusses itemspezifischer und personenspezifischer Prädiktoren auf das Antwortverhalten erlauben. Die Auswertung der Antworten von N=2112 Versuchspersonen auf 12 Items mit einem sechsstufigen Antwortformat zum Persönlichkeitskonstrukt *Personal Need for Structure* zeigt, dass ein mehrdimensionales IRT-Modell zur Berücksichtigung von mittleren und extremen Antworttendenzen einem eindimensionalen Modell überlegen ist. Die Validität der latenten Dimensionen als Reflektion des inhaltlichen Persönlichkeitskonstrukts einerseits und von Antworttendenzen andererseits wird durch differentielle Korrelationen zwischen den latenten Dimensionen, durch gegenläufige Effekte itemspezifischer Prädiktoren der Itemposition sowie durch selektive und dissoziierende Effekte personenspezifischer Prädiktoren wie Extraversion, Offenheit und Gewissenhaftigkeit demonstriert.

Böckenholt, U. (under review). Modeling multiple response processes in judgment and choice.

De Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. Journal of Marketing Research, 45, 104-115.

Rijmen, F., Tuerlinckx, F., de Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. Psychological Methods, 8, 185-205.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. Psychological Methods, 15, 96-110.

Conceptual difficulties of conventional ways of modeling mediation

Andreas Klein
Holger Brandt

J.W. Goethe-Universität Frankfurt
Institut für Psychologie,
Mertonstr. 17
60054 Frankfurt a.M.
Campus Bockenheim

aklein2569@gmail.com

In many research contexts in psychology, the social or health sciences, researchers are interested in identifying intermediate variables that stand in the pathway or may serve as early indicators of a developing disease or critical behavior. Causal effects that are indirect (mediated) and go through intermediate variables that act as components in a causal chain are also known as mediator effects. In the biostatistical literature, however, very serious concerns about the conventional methodology such as path analysis ("Kenny's Mediator Model") used to explore those effects have been raised. Following this critique, a causal interpretation of a path-analytic mediator effect is hardly ever justified, and it can be shown that path analysis may often identify supposedly strong mediator effects that in reality are nothing but methodological artefacts. On the other hand, path analysis has seen routine application as a standard technique to analyze potential mediators in the psychological literature. In this paper, we explain why the standard mediator model is conceptually flawed, and why it requires additional, untestable assumptions to become interpretable as a model that can represent a mediated causal process. Directions for future model developoment and consequences for practical applications are discussed.

Identification of mediator variables in randomized longitudinal trials

Holger Brandt
Andreas Klein

Abteilung für Psychologische Methodenlehre Evaluation und Forschungsmethodik Goethe-Universität Frankfurt
Mertonstr. 17
60054 Frankfurt

brandt@psy.uni-frankfurt.de

In intervention studies concerning treatment efficacy the interest in mediator variables has been increasing. Mediator variables are variables which act between an intervention and an outcome variable and which are intended to capture the effect of the intervention on the outcome. The knowledge about mediator variables enhances the understanding of the mechanisms transmitting the effect of the intervention. Different approaches have been suggested in order to identify potential mediator variables, which have in common that they condition on the mediator variable in order to quantify the mediated effect (e.g. Baron & Kenny, 1986; Prentice, 1989). However, serious concerns about this type of methodology have been raised in the more recent statistical literature.

In this paper, we explain in more detail what steps might be taken to modify the conventional approach, and we propose a method that helps to reduce the bias for the estimation of the mediated effect, when the data structure meets specific requirements. The differences between the conventional and the proposed method are exemplified by a small simulation study and an empirical example. The results indicate that in real-world situations the conventional mediator model mostly leads to a strongly biased estimation of the mediated effect. In comparison, the proposed method leads to a reduced bias.

Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173-1182.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. Statistics in Medicine, 8, 431-440.

Die Kombination von Social Relations- und Multilevel Mediation Analysen im Klassenkontext

Katrin Rentzsch

Lehrstuhl Persönlichkeitspsychologie und Psychologische Diagnostik
Universität Bamberg
Markusplatz 3
96047 Bamberg

katrin.rentzsch@uni-bamberg.de

Interpersonelle Wahrnehmungen im Klassenkontext sind dyadisch und damit komplexer als oftmals angenommen. Die Einschätzung eines Mitschülers hängt somit nicht nur von Eigenschaften des beurteilten Schülers ab, sondern auch von Eigenschaften des Beurteilers. So zum Beispiel hängt Bernds Einschätzung von Anna als „Streberin“ nicht nur von Annas Charakteristika ab, sondern auch von Bernds Tendenz, Mitschüler zu mobben. Das Social Relations Modell (SRM; Kenny, 1994) gibt für solche interpersonellen Wahrnehmungen bzw. Verhaltensweisen die notwendigen inhaltlichen und statistischen Implikationen. Das SRM besagt, dass die Einschätzung einer anderen Person in drei Komponenten aufgeteilt werden kann: den Perceiver Effekt (d.h. die durchschnittliche Tendenz des Beurteilers, andere auf bestimmte Weise wahrzunehmen), den Target Effekt (d.h. die durchschnittliche Tendenz des Beurteilten, von anderen auf bestimmte Weise wahrgenommen zu werden) und den Relationship Effekt (d.h. die spezifische Tendenz des Beurteilers, einen spezifischen Beurteilten auf bestimmte Weise einzuschätzen). Mittels Social Relations Analysen können aus den Einschätzungen von Klassenkameraden die Social Relations Effekte (d.h. Perceiver-, Target-, und Relationship Effekte) extrahiert und in weiterführende Analysen zur Hypothesentestung integriert werden. Zur Untersuchung von Prozessen beispielsweise können die Social Relations Effekte mittels Mediationsanalysen weiterverrechnet werden. Um der hierarchischen Struktur des Datensatzes gerecht zu werden, empfehlen sich hier sogenannte Multilevel Mediation Analysen. In dem Vortrag werden die dem Social Relations Modell zugrundeliegenden Erfassungsmethoden vorgestellt, die Datenanalyseverfahren und -Software anhand einer Beispielstudie verdeutlicht und der praktische Einsatz bei Untersuchungen im Klassenkontext diskutiert.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York, NY: Guilford Press.

Analyse individueller Klassifikationen in verallgemeinerten DINA-Modellen

Ann Cathrice George[1]
Alexander Robitzsch[2]
Jürgen Groß[3]

[1]Research School Education and Capabilities
TU Dortmund
Hauert 14a
44227 Dortmund,
Deutschland

a.george@educap.de

[2]Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens (BIFIE)
Salzburg
Zentrum für Bildungsmonitoring & Bildungsstandards Alpenstraße 121
5020 Salzburg
Österreich
a.robitzsch@bifie.at

[3]Fakultät Statistik, TU Dortmund
Vogelpothsweg 87
44227 Dortmund
Deutschland

gross@statistik.tu-dortmund.de

Kognitive Diagnostische Modelle (Cognitive Diagnosis Models; CDMs; Di Bello, Roussos & Stout, 2007) bilden eine Familie konfirmatorischer probabilistischer Modelle mit dichotomen latenten Variablen, die Mehrfachladungsstrukturen erlauben. Häufig wird der Einsatz dieser Modellfamilie damit begründet, dass individuelle Klassifikationen der Testteilnehmer in latente dichotome Klassen von beherrschten bzw. nicht beherrschten Grundfähigkeiten nutzbringend zur Erstellung informativer Rückmeldungen eingesetzt werden können.

Im vorliegenden Beitrag werden verschieden komplexe Modelle der Klasse der verallgemeinerten DINA-Modelle (G-DINA; de la Torre, 2011; Henson, Templin & Willse, 2009) für den Datensatz zur österreichischen Bildungsstandards-Testung der 8. Schulstufe in Mathematik spezifiziert und auf Eigenschaften untersucht. Dabei wird zunächst die Modellanpassung mit Hilfe verschiedener Fitindizes bewertet. Außerdem werden die unter den verschiedenen Modellen entstehenden Klassifikationen der einzelnen SchülerInnen als KönnerInnen bzw. NichtkönnerInnen der modellierten Fähigkeiten verglichen. Besonders hervorgehoben wird dabei die individuelle Klassifikation hinsichtlich der inhaltlichen Kompetenzen, die in jedem der Modelle spezifiziert sind.

In einer sich anschließenden Simulationsstudie werden die Unterschiede zwischen dem gewöhnlichen DINA Modell und dem für die Daten komplexesten G-DINA Modell bewertet. Da i.A. nicht bekannt ist aus welchem Modell empirische Testdaten stammen, werden in der Simulationsstudie die Daten aus einem der beiden Modelle generiert und mit dem anderen angepasst. Die Größe der jeweils entstehenden Fehlklassifikationsraten soll Hinweise liefern, in welchen Situationen ein gewöhnliches oder ein komplexes Modell vorzuziehen ist.

de la Torre, J. (2011). The generalized DINA framework. Psychometrika, 76, 179-199.

DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao, & S. Sinharay (Eds.), Handbook of statistics (pp. 979–1030). Amsterdam: Elsevier.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. Psychometrika, 74, 191-210.

Clusteranalyse oder Latent-Class-Analyse? Ein Vergleich der Verfahren am Beispiel von Alkoholkonsummustern Jugendlicher in Europa

Astrid-Britta Bräker
Stefan Huber
Renate Soellner

Universität Hildesheim
Institut für Psychologie
Marienburger Platz 22
31141 Hildesheim

braeker@uni-hildesheim.-de

Hintergrund: Jugendliche trinken heute nicht nur früher Alkohol, ihr Konsum nimmt auch immer exzessivere Ausmaße an. Im Rahmen des von der Europäischen Kommission finanzierten Projektes Alcohol Abuse among Adolescents in Europe (AAA-Prevent) wird eine Teilstichprobe der International Self-Report Study of Delinquency (ISRD-2) von 57.771 zwölf- bis sechszehnjährigen Jugendlichen aus 25 europäischen Ländern erneut analysiert. Ziel ist es, den Alkoholkonsum der Siebt- bis Neuntklässler über die Länder hinweg zu beschreiben, Faktoren für riskanten Konsum sowohl auf individueller als auch auf Länderebene zu identifizieren und deren Einfluss auf den Alkoholkonsum zwischen den Ländern zu vergleichen, um schließlich effektive Präventionsstrategien zu empfehlen. Um nicht über einzelne Konsumvariablen, sondern über Konsummuster Aussagen treffen zu können, sollen Konsumentenprofile aufgedeckt werden.

Methode: Zur Identifikation von Konsummustern wurden einerseits Clusteranalysen (Ward, k-means), andererseits latente Profil-Analysen als Variante der Latent-Class-Analyse durchgeführt. Hierbei wurden Angaben hinsichtlich der Häufigkeit und Menge des Bier-, Wein- und Spirituosenkonsums von insgesamt N = 27.653 Jugendlichen im Alter von 12 bis 16 Jahren einbezogen, die in ihrem Leben bereits Alkohol getrunken haben. Beide Methoden und ihre Befunde werden beschrieben, miteinander verglichen und diskutiert.

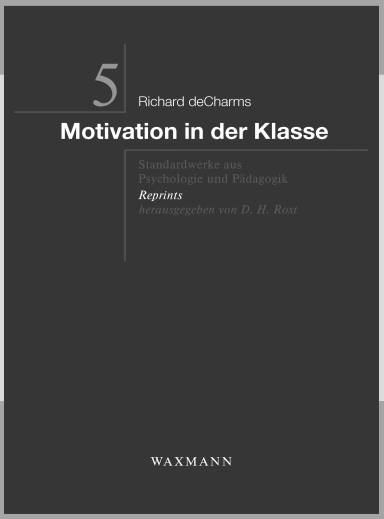
Ergebnisse: Mithilfe von Clusteranalysen in SPSS 18.0 konnten folgende vier Konsumentengruppen identifiziert werden: 1. sehr selten und sehr wenig Konsumierende (73.6%), 2. moderat häufig und moderat viel Konsumierende (19.9%), 3. moderat häufig und sehr viel Konsumierende (3.8%) und 4. sehr häufig und moderat viel Konsumierende (2.7%). Diese Lösung wird verglichen mit den Ergebnissen unterschiedlich restriktiver konfirmatorischer Latent-Profile Analysen, die mit MPlus 5.1 ermittelt wurden. Hier kann die Lösung der CA bei vorgegebenen Antwortmustern repliziert werden und stellt mit einem Entropy-Wert von .90 eine zufriedenstellende Klassifikation dar, obwohl der explorative Vergleich verschiedener Modelle ohne Parameter-Vorgaben eine Sechs-Klassen-Lösung nahe legt. Diese Lösung ist jedoch schwer interpretierbar und zeigt deutlich ungleich verteilte Klassengrößen (0.2 – 89%).

Diskussion: Trotz der aktuellen Popularität der latenten Klassen- bzw. Profile-Analyse scheint das explorative Verfahren der Clusteranalyse für die hier angestrebte Fragestellung eine inhaltlich besser interpretierbare Lösung bereit zu stellen. Implikationen für die weitere Anwendung der Methoden werden besprochen.

Personenverzeichnis

Standardwerke aus Psychologie und Pädagogik – Reprints

herausgegeben von Detlef H. Rost



Band 5

Richard deCharms

Motivation in der Klasse

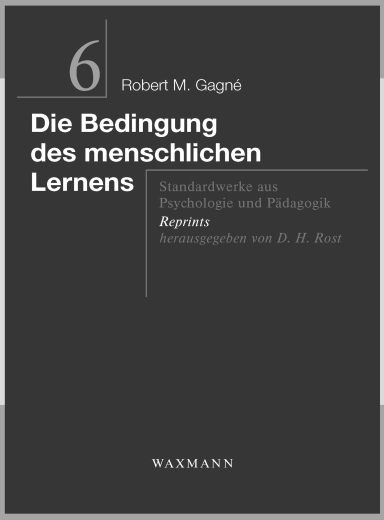
2011, 250 Seiten, br., 29,90 €, ISBN 978-3-8309-2583-5

Band 6

Robert M. Gagné

Die Bedingung des menschlichen Lernens

2011, 312 Seiten, br., 34,90 €, ISBN 978-3-8309-2584-2



WAXMANN
Münster ▪ New York ▪ München ▪ Berlin