Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FB Psychologie
Psychologische Methoden

Discussion Paper

No. 2

Fundamentale Probleme beim Einsatz testtheoretischer Modelle zur Diagnose von Individuen

July 2013

Pascal Jordan & Martin Spiess

Psychologische Methoden, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany

Fundamentale Probleme beim Einsatz testtheoretischer Modelle zur Diagnose von Individuen

Pascal Jordan* und Martin Spieß

Universität Hamburg

*Korrespondenz: Von Melle-Park 5, 20146 Hamburg; e-Mail: pascal.jordan@uni-hamburg.de.

Zusammenfassung

Ziel dieses Beitrags ist es, mit Bezug auf ein erst kürzlich „entdecktes“ Paradoxon, den Wert der gängigen testtheoretischen Modelle für diagnostische Zwecke zu hinterfragen. Nachdem zuerst dargelegt wird, dass der in der Praxis etablierte Schätzer für die Fähigkeit einer Person - der Summenscore - aus statistischer Sicht unzureichende Eigenschaften aufweist, erfolgt eine Betrachtung des modellbasierten Schätzers. Obwohl dieser Schätzer aus statistischer Perspektive optimal ist, erweist er sich in Hinblick auf die Fairness einer Diagnose als inakzeptabel. Schätzer, die dem testtheoretische Modell inhärent sind, weisen somit für die Individualdiagnostik gravierende Mängel auf. Als logische Konsequenz stellt dies zugleich den Wert des testtheoretischen Modells in Frage.

1 Einleitung

Das Konzept einer latenten Variable ist ein integraler Bestandteil der psychologischen Forschung. Konstrukte wie „Intelligenz“, „Extraversion“ oder „Leistungsmotivation“ stellen Beispiele von für die psychologische Theorie zentralen Begriffen dar, die per Definition nicht direkt beobachtbar sind. Anhand einer geeigneten Auswahl von Items ist es jedoch möglich, Rückschlüsse auf diese latenten Konstrukte vorzunehmen. Grundlage hierfür bietet ein stochastisches Modell, das den Ausdruck: „Ein Item misst eine latente Größe“ formalisiert und zugleich eine empirische Überprüfung der Messqualitäten des Messinstruments (d.h. des Fragebogens) ermöglicht.

Die statistischen Modelle lassen sich auf verschiedene Weise klassifizieren. Neben der gängigen Unterscheidung zwischen Modellen der klassischen Testtheorie (bzw. faktoranalytischen Modellen) und Modellen der probabilistischen Testtheorie stellt die Einordnung gemäß der Dimensionalität des latenten Konstrukts einen häufig diskutierten und zentralen Aspekt dar. Eindimensionale Tests, d.h. Tests, die genau ein Konstrukt messen, gewährleisten sowohl eine eindeutige Interpretation eines Testergebnisses als auch den adäquaten Vergleich der Testwerte zweier Personen (McNemar, 1946, S.298). Diesen theoretischen Vorteilen eindimensionaler Tests stehen jedoch diverse praktische bzw. operationale Nachteile gegenüber. So ist aus zahlreichen Quellen (z.B. Rosenbaum, 1984; Holland und Rosenbaum, 1986; Salguiero u.a., 2008) bekannt, dass eindimensionale Modelle so starke Restriktionen an die Daten stellen, dass sie in der Anwendung praktisch nicht vorkommen. Im Kontext des faktorenanalytische Modells beispielsweise zeigt Shapiro (1982) die Unmöglichkeit auf, einen *eindimensionalen* Test mit ausreichender Länge zu konstruieren, und Ten Berge und Socan (2004) weisen zusätzlich auf das antagonistische Verhältnis zwischen der Reliabilität einer Skala und der Forderung nach ihrer Eindimensionalität hin.

Neben diesen formalen Argumenten lassen sich aber in vielen Fällen auch sachlogische Gründe anführen, warum eine Skala nicht eindimensional sein kann. So erfordert z.B. ein

Physiktest außer mathematischen Fähigkeiten auch räumliches Vorstellungsvermögen, Textaufgaben mit mathematischem Inhalt setzen mathematische- sowie Lesefähigkeiten voraus und ein Fragebogen zur Messung von Extraversion beinhaltet mitunter auch Items, die „Sensation-Seeking“ miterfassen. Aufgrund dieser höheren Realitätsnähe stellen mehrdimensionale Tests in der Praxis die Regel und eindimensionale Tests die Ausnahme dar[1]. Wir können im Folgenden daher ohne Beschränkung der Allgemeinheit davon ausgehen, dass ein in Wahrheit mehrdimensionaler Test vorliegt - ohne damit auszuschließen, dass dieser vom Anwender als eindimensional „behandelt“ wird. Bezüglich dieser Ausgangssituation werden wir zunächst (Abschnitt 2) aufzeigen, dass die in der Praxis übliche Kennzeichnung der zu diagnostizierenden Person - nämlich die Angabe von Summenwerten auf (nach einer bestimmten Regel gebildeten) Subskalen - aus statistischer Sicht fragwürdig ist. Die anschließende Betrachtung (Abschnitt 3) des modellbasierten, statistisch effizienten Schätzers wird zwar dessen stochastische Überlegenheit demonstrieren, jedoch zugleich herausstellen, dass dieser Schätzer einer Paradoxie unterliegt. Dieser paradoxe Effekt, der sich in einem Leistungstest z.B. dahingehend manifestieren kann, dass korrekte Antworten schädlich für die Person im Hinblick auf eine positive Diagnose sind, besitzt weitreichende Implikationen für die Praxis und wird zunächst in Abschnitt 4 in allgemeiner Form diskutiert, bevor seine Konsequenzen anhand eines realen Datensatzes (Abschnitt 5) verdeutlicht werden.
Ausgehend von diesen Beobachtungen (statistische Effizienz und Paradoxie des modellbasierten Schätzers) ziehen wir im abschließenden sechsten Abschnitt allgemeine Schlussfolgerungen über den Einsatz testtheoretischer Modelle in der psychologischen Diagnostik.

[1]Explorative Faktorenanalysen weisen auch nach Varimax-Rotation noch beträchtliche Querladungen auf und konfirmatorische Faktorenanalysen lehnen zumeist die Nullhypothese der Eindimensionalität ab.

2 Der Summenscore als Schätzer für die Fähigkeit einer Person

Wir beschränken uns im Folgenden auf faktorenanalytische Modelle[2], da diese gegenüber Modellen der probabilistischen Testtheorie in der Praxis wesentlich stärker verbreitet sind. Da wir lediglich am individualdiagnostischen Einsatz der faktorenanalytisch konstruierten Skala interessiert sind, nehmen wir zudem an, dass die Eigenschaften der Items (z.B. die Ladungen auf den latenten Dimensionen) anhand einer hinreichend großen Kalibrierungs/Normierungsstichprobe geschätzt wurden und somit als näherungsweise bekannt betrachtet werden können. Wir gehen also davon aus, dass die Charakteristiken der Messinstrumente (der Items) bekannt sind und dass somit die einzige unbekannte Größe die (mehrdimensionale) Fähigkeit[3] der zu diagnostizierenden Person ist.

Nach der „konventionellen“ Methode wird die zu diagnostizierende Person durch Summenwerte auf einzelnen Subskalen gekennzeichnet. Die Bildung der Subskalen beruht dabei i.d.R auf folgendem Prinzip: Jedes Item wird basierend auf seinen Ladungen exakt einer Dimension zugeordnet, d.h. die dem Item inhärente Information fliesst in (nur) eine Dimension ein. Zudem folgt anhand der Verrechnungsvorschrift (Bildung des *ungewichteten* Summenwerts), dass alle Items innerhalb einer derart gebildeten Subskala einen gleichwertigen Beitrag zur Kennzeichnung der Person liefern.

Zwei offensichtliche Einwände für diese Vorgehensweise ergeben sich: Erstens ist eine Zuordnung eines Items zu *einer* Dimension offensichtlich nur gerechtfertigt, wenn die Ladungen auf den anderen Dimensionen verschwinden, d.h. wenn von Null abweichende Ladungen auf anderen Dimensionen nur aufgrund des Stichprobenfehlers vorhanden sind. Zweitens ist die Austauschbarkeit der Items innerhalb einer Subskala bezüglich ihres Beitrags zur Kennzeich-

[2]Eine analoge Diskussion im Rahmen der probabilistischen Testtheorie wäre leicht durchführbar; wird jedoch aus Platzgründen unterlassen.

[3]Wir benutzen den Begriff „Fähigkeit“ lediglich zur sprachlichen Vereinfachung. Alle folgenden Ergebnisse sind ebenso für allgemeine „latente Dimensionen“ (z.B. Depression, Extraversion, ...) gültig.

nung der Person nur dann sinnvoll, wenn die Items „gleich gute“ Messinstrumente darstellen (in Termini der Klassischen Testtheorie: wenn es sich um parallele Messungen handelt).
Die Angemessenheit des Summenwerts zur Kennzeichnung der Person beruht folglich implizit auf zwei Prämissen: Der Einfachstruktur der Ladungsmatrix (d.h. jedes Item lädt auf genau einer Dimension) sowie der Parallelität der Items innerhalb einer Subskala. Sowohl auf formaler Ebene (siehe z.B. Shapiro, 1982; Ten Berge und Socan, 2004) als auch auf sachlogischer Ebene (Ist es plausibel, dass die Bearbeitung eines Items nur *eine* Fähigkeit erfordert?) lassen sich Argumente anführen, die darauf hinweisen, dass beide Forderungen unrealistisch sind und in der Praxis nahezu nie erfüllt sein werden.
In der Tat liesse sich ein faktorenanalytisches Modell, welches die beiden Forderungen erfüllt, bezüglich seiner empirischen Restriktivität mit einem Rasch-Modell vergleichen[4]. Die Schwierigkeit, einen Rasch-skalierbaren Test zu konstruieren, ist wiederum wohldokumentiert.
Bevor wir zu der Diskussion eines statistisch „angemessenen“ Schätzers übergehen, sollen zur Verdeutlichung die statistischen Eigenschaften des Summenwerts als Schätzer für die Fähigkeit der Person aufgeführt werden, wenn die obigen Annahmen (Einfachstruktur und Parallelität) verletzt sind[5]:

- Der Summenwert ist i.A. kein unverzerrter Schätzer. Die Verzerrtheit kann (in Abhängigkeit von der Ladungsmatrix) sogar mit steigender Testlänge zunehmen. Mit anderen Worten: Das Messen mit weiteren Messinstrumenten (Items) kann bezüglich der Unverzerrtheit des Schätzers schädlich sein.

- Der Summenwert ist i.A. kein konsistenter Schätzer, d.h. das zusätzliche administrieren von weiteren Items impliziert nicht, dass die Schätzung dem wahren Wert (in Wahrscheinlichkeit) näher kommt.

[4]Man vergleiche z.B. die faktorenanalytischen Ergebnisse von Salguiero u.a. (2008) mit den IRT- basierten Ergebnissen von Rosenbaum (1984) sowie die Aussagen zur Suffizienz des Summenscores (Fischer, 1995; Lehmann und Casella, 1998, S.176ff.) in den beiden Modellen.
[5]Beweise dieser Eigenschaften können von den Autoren angefordert werden.

- Der Summenwert ist kein suffizienter Schätzer, d.h. das Antwortmuster der Person enthält zusätzliche Information über die Fähigkeit einer Person. Zwei Personen mit gleichem Summenwert sollten nicht gleichermaßen diagnostiziert werden.

Diese (nicht erschöpfende) Aufzählung zeigt, dass die in der Praxis übliche Kennzeichnung einer Person durch Subskalen-spezifische Summenwerte vor dem Hintergrund des testtheoretischen Modells fragwürdig ist.

3 Der modellbasierte Schätzer der Fähigkeit einer Person

Nachdem skizziert wurde, warum der einfache Summenwert i.d.R. keine adäquate Kennzeichnung der Person gewährleistet, sollen in diesem Abschnitt Eigenschaften des modellbasierten Schätzers dargelegt werden. Es sei dabei betont, dass dieser Schätzer (sog. Bartlett-Score) optimal ist - wenn man das Modell, welches der Konstruktion des Tests zugrundeliegt, „ernst nimmt“. Er besitzt u.a. die folgenden Eigenschaften[6]:

- Der Bartlett-Score ist ein unverzerrter Schätzer. Dies gilt für jede beliebige Testlänge.
- Unter allen linearen unverzerrten Schätzern stellt der Barlett-Score den Schätzer mit kleinster Varianz dar.
- Er ist (unter schwachen Voraussetzungen an die Ladungsmatrix) konsistent.
- Der Bartlett-Score stellt eine erschöpfende Beschreibung der Person dar, d.h. Kenntnis

[6]Berücksichtigt man die Analogie zum gewichteten Kleinste-Quadrate-Schätzer im linearen Modell, so lassen sich sämtliche aufgeführte Eigenschaften aus den entsprechenden Resultaten über den gewichteten KQ-Schätzer herleiten (siehe z.B. Mardia u.a. 1979, S. 171ff.; Lehmann und Casella, 1998, S.130).

des Bartlett-Scores stellt gleichwertige Information gegenüber Kenntnis des gesamten Antwortmusters dar.

Man beachte, dass der Bartlett-Score zudem den Maximum-Likelihood-Schätzer unter dem postulierten testtheoretischen Modell (Faktorenanalyse mit „üblicher“ Annahme einer Normalverteilung) darstellt. Insofern lässt er sich als modellinhärenter Schätzer betrachten.
Im Gegensatz zu dieser stochastischen Optimalität werden wir im folgenden Abschnitt seine Unangemessenheit für individualdiagnostische Zwecke darstellen. Dies wird zugleich den zentralen Konflikt zwischen statistischer Effizienz und (Einzelfall-bezogener) Fairness einführen.

4 Paradoxe Effekte bei der Verwendung modellbasierter Schätzers

Ziel dieses Abschnitts[7] ist es, auf einen fundamentalen „Mangel“ (mehrdimensionaler) modellbasierter Schätzer hinsichtlich der Diagnose von Individuen hinzuweisen, der erst 2009 von Hooker, Finkelman und Schwartzman aufgezeigt wurde. Dieser „Mangel“ manifestiert sich in Gestalt eines Paradoxons bei der Fähigkeitsschätzung einer Person. Besonders schwerwiegende Konsequenzen ergeben sich bei der Verwendung dieser Fähigkeitsschätzung zur Diagnose von Personen - etwa mit Hilfe von festgelegten Schwellenwerten, die zu Überschreiten sind, damit eine entsprechend positive Klassifikation erfolgt. Auch wenn wir uns im Wesentlichen auf den Schätzer im Rahmen des faktoranalytischen Modells (Bartlett Score) beschränken, sei darauf hingewiesen, dass das folgende Paradoxon weit über dieses Modell hinausgeht. Es zeigt sich in nahezu jedem mehrdimensionalen testtheoretischen Modell unabhängig davon, ob es sich um ein Modell der klassischen oder um ein Modell der

[7]Die mathematischen Beweise der relevanten Aussagen dieses Abschnitts finden sich für dichotome probabilistische Modelle bei Hooker u.a. (2009) und für ordinale bzw. faktorenanalytische Modelle bei Jordan und Spiess (2011).

probabilistischen Testtheorie handelt.

Die grundlegende inhaltliche Aussage des Paradoxons kann am besten anhand des folgenden Beispiels von Hooker u.a. (2009) eingeführt werden:

„Jane and Jill are fast friends who are nonetheless intensely competitive. At the end of high school they each take an entrance exam for a prestigious university. After the exam, they compare notes and discover that they gave the same answers for every question but the last. On checking their materials, it is clear that Jane answered this question correctly, but Jill answered incorrectly. They are therefore very surprised, when the test results are published, to find that Jill passed but Jane did not!“ (Hooker u.a., 2009, S.419)

Auch wenn dieses Resultat zunächst befremdlich erscheint, so lässt es sich dennoch mathematisch fundieren. Das wesentliche Prinzip des Paradoxons beruht auf einer kontraintuitiven Klassifikation zweier Personen. Es ist möglich, dass eine Person die Fragen besser bearbeitet als eine andere Person und dennoch eine schlechtere Klassifikation erhält. Eine selten explizit ausgedrückte Forderung an einen Test liegt in der Annahme begründet, dass „bessere“ Antworten zu einer „besseren“ Klassifikation führen. Zumindest sollten richtige Antworten eine Person nicht schlechter stellen.

Oder etwas formaler ausgedrückt: Besteht bezüglich der Antwortvektoren $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ zweier Personen eine Ordnung, d.h. wenn für alle Items i gilt $x_{2i} \geq x_{1i}$, so soll die Ordnung der Personen auch in den geschätzten Fähigkeiten erhalten bleiben, d.h. es soll unter der Annahme, dass keine Fähigkeit negativ zur Lösung beiträgt, gelten: $\hat{\theta}_{2l} \geq \hat{\theta}_{1l}$ ($\hat{\theta}_{2l}$ bezeichnet hierbei den Schätzwert, den die zweite Person auf der l-ten Dimension erhält) für alle Dimensionen l. Die (partielle) Ordnung von Personen anhand ihrer Antworten soll somit auch in den Schätzungen der Fähigkeiten weiter bestehen bleiben. Jede Form der Fähigkeitsschätzung, die dieses Invarianzprinzip verletzt, scheint inakzeptabel. Es ist genau die Verletzung dieses Invarianzprinzips bei mehrdimensionalen Tests, auf die das obige Beispiel abzielt. Eine heuristische Erklärung des paradoxen Effekts lässt sich am besten anhand eines Tests der

zwei latente Dimensionen misst - z.B. mathematische Fähigkeit und räumliches Vorstellungsvermögen - geben.

Angenommen der Test besteht aus zehn Items, wobei die ersten neun Items von beiden Personen - Jane und Jill - identisch bearbeitet wurden und gesetzt den Fall, das zehnte, unterschiedlich beantwortete Item misst die mathematische Fähigkeit aber nicht die räumliche Fähigkeit[8], dann lässt sich informell folgendermaßen argumentieren:

Die beiden Personen weisen nach dem neunten Item identische Fähigkeitsschätzungen auf. Im Gegensatz zu Jill löst Jane das zehnte Item, folglich ist ihre mathematische Fähigkeit (vermutlich) ausgeprägter. Wenn wir für den Moment diesen Unterschied in der mathematischen Fähigkeit als Fakt annehmen, dann stellt sich die Frage, wie dieser Aspekt mit der identischen Performance bei den ersten neun Items in Beziehung steht. Mit dem Wissen, dass Janes mathematische Fähigkeit ausgeprägter ist und mit dem zusätzlichem Wissen, dass beide Fähigkeiten positiv zur Lösung der ersten neun Items beitragen, ist der Rückschluss auf ein geringeres räumliches Vorstellungsvermögen plausibel. Wenn Jane nämlich über eine ausgeprägtere mathematische Fähigkeit *und* mindestens so gutes räumliches Vorstellungsvermögen wie Jill verfügen würde, dann wäre eine identische Performance der beiden Personen bei den ersten neun Items nicht plausibel. Folglich schließen wir auf ein geringeres räumliches Vorstellungsvermögen von Jane gegenüber Jill. Wenn gleich diese Schlussfolgerung auf einigen Vereinfachungen beruht (z.B.: bekannter Fähigkeitsunterschied in der Mathematik-Dimension), unterscheidet sie sich nur in einem „technischen Detail“ von der vollständigen Argumentation des entsprechenden allgemeingültigen mathematischen Beweises.

Die dargelegte Argumentation stützt sich im Wesentlichen auf die Attribution eines Testergebnisses auf zwei (bzw. mehrere) latente Größen. Sie lässt sich nicht für eindimensionale Modelle durchführen. In eindimensionalen Modellen (z.B. dem Rasch-Modell) führt ein besseres Antwortmuster auch zu einer höheren geschätzten Fähigkeit. In zweidimensionalen

[8]Die Annahme eines „reinen“ Mathematikitems dient lediglich zur Vereinfachung der folgenden Argumentation. Die Herleitung des paradoxen Effekts ist nicht daran gebunden.

Modellen ist es hingegen möglich, dass eine der beiden Dimensionen durch eine richtige Antwort geringer geschätzt wird. Dies stellt insbesondere bei der anschließenden Verwendung von (zwei) Schwellenwerten, deren Überschreiten erforderlich ist, um eine bestimmte positive Klassifikation zu erlangen, ein Problem dar. Es kann vorkommen, dass eine Person durch eine falsche Antwort an Stelle einer richtigen Antwort auf der zweiten Dimension höher geschätzt wird. Dieser Anstieg kann zu einem Überschreiten des Schwellenwerts bezüglich dieser Dimension führen. Wenn zugleich der Schwellenwert der ersten Dimension durch diese falsche Antwort nicht unterschritten wird, resultiert folglich eine positive Klassifikation, die durch eine falsche Antwort verursacht wurde.

Dass das obige Szenario kein rein akademisches Resultat ist, lässt sich anhand der Untersuchung der Prävalenz dieses Effekts in realen Datensätzen verdeutlichen. So berichten Finkelman u.a. (2010) beispielsweise von einem (probabilistischen) Test, in dem für 15.4% der Personen eine bessere Klassifikation durch das Geben von ein oder mehreren Falsch-Antworten möglich gewesen wäre. Unabhängig von diesem empirischen Resultat lässt sich im Rahmen des faktorenanalytischen Modells das erstaunliche Resultat herleiten, dass jedes Individuum (d.h. jedes individuelle Antwortmuster) von dem paradoxen Effekt betroffen ist, d.h. es existiert stets ein Item, bei dem eine falsche Antwort (allgemeiner: eine Verringerung des erzielten Scores) die Schätzung der Fähigkeit in einer Dimension (z.B. räumliches Vorstellungsvermögen) erhöht, obwohl diese Fähigkeit positiv zur Lösung des Items beiträgt.

Bevor wir den paradoxen Effekts in der KTT anhand eines realen Datensatzes demonstrieren, möchten wir an dieser Stelle nochmals die drei wesentlichen Aussagen hervorheben:

Das Paradoxon tritt nur bei mehrdimensionalen Tests auf (Mehrdimensionalität); es führt zu einem unfairen und kontraintuitiven Vergleich zweier Personen (Inter-Personen-Vergleich); (absichtliches) Falsch-Antworten kann sich für eine Person als nützlich erweisen (Intra-Personen-Vergleich).

5 Paradoxe Effekte des Bartlett-Scores

Dieser Abschnitt soll anhand eines konkreten Zahlenbeispiels den paradoxen Effekt im Rahmen des faktoranalytischen Modells demonstrieren. Bevor wir den Datensatz vorstellen, geben wir eine kurze und allgemeine Beschreibung des zugrundegelegten testtheoretischen Modells. Dies wird zugleich die Einführung der relevanten Notation ermöglichen und somit die Grundlage für die weitergehenden Erläuterungen des Bartlett-Scores bilden.

5.1 Theoretische Grundlagen

Der Ausgangspunkt der folgenden Analysen ist das faktorenanalytische Modell mit der „üblichen“ Normalverteilungsannahme (siehe z.B. Mardia u.a., 1979, Kap. 9). Durch eine entsprechend große Kalibrierungsstichprobe seien die Itemcharakteristiken, d.h. die Einträge der $(k \times p)$ Ladungsmatrix Λ und die Fehlervarianzen $(\psi_i)_{i=1\ldots k}$ (ψ_i entspricht der Fehlervarianz des i-ten Items), hinreichend genau geschätzt und somit (näherungsweise) bekannt. Die einzige unbekannte Größe bei der Diagnose einer Person ist dann die Ausprägung des (potentiell mehrdimensionalen) latenten Konstrukts $\boldsymbol{\theta}$. Auf diese Ausprägung kann anhand des beobachteten Antwortmusters $\boldsymbol{x}$ (ein Vektor, dessen j-ter Eintrag den auf dem j-ten Item erzielten Score widerspiegelt), das dem Modell $\boldsymbol{x} \sim N(\Lambda\boldsymbol{\theta}, \Psi)$ folgt, geschlossen werden. Ψ bezeichnet eine Diagonalmatrix bestehend aus den Fehlervarianzen $(\psi_i)_{i=1,\ldots k}$.
Der modellbasierte Schätzer (Bartlett-Score) entspricht in diesem Kontext dem gewichteten Kleinste-Quadrate-Schätzer. Er ist formal darstellbar als:

$$\hat{\boldsymbol{\theta}} := (\Lambda^T \Psi^{-1} \Lambda)^{-1} \Lambda^T \Psi^{-1} \boldsymbol{x}. \tag{1}$$

Im Folgenden setzen wir voraus, dass die Ladungsmatrix keine negativen Einträge besitzt. Dies soll den Sachverhalt wiedergeben, dass alle Dimensionen (z.B. mathematische Fähigkei-

ten und räumliches Vorstellungsvermögen) positiv zur Beantwortung der Items beitragen. Eine Erhöhung des Wertes auf einer latenten Dimension erhöht somit den zu erwartenden Score auf dem betrachteten Item[9]. Da jede Dimension einen positiven Beitrag leistet, wäre zu erwarten, dass eine bessere Performance einer Person in dem Test auch zu einer höheren geschätzten Fähigkeit bezüglich jeder Dimension führt.

Wir schreiben im Folgenden $\boldsymbol{x_2} > \boldsymbol{x_1}$, wenn Person 2 auf jedem Item mindestens den entsprechenden Score von Person 1 erzielt hat und wenn zudem (mindestens) ein Item existiert, bei dem Person 2 einen höheren Score erzielt hat. Für zwei Antwortmuster $\boldsymbol{x_2}$ und $\boldsymbol{x_1}$ resultiert gemäß (1) als Differenz der geschätzten Fähigkeiten der folgende Ausdruck:

$$\hat{\boldsymbol{\theta}}_2 - \hat{\boldsymbol{\theta}}_1 := (\Lambda^T \Psi^{-1} \Lambda)^{-1} \Lambda^T \Psi^{-1} (\boldsymbol{x}_2 - \boldsymbol{x}_1). \tag{2}$$

Ein paradoxer Effekt tritt auf, wenn sich eine bessere Performance $\boldsymbol{x_2} > \boldsymbol{x_1}$ nicht in den Fähigkeitskomponenten widerspiegelt, d.h. wenn mindestens eine Dimension l existiert, für die $\hat{\theta}_{2l} < \hat{\theta}_{1l}$ gilt. In diesem Fall wirkt sich ein besseres Abschneiden in dem Test negativ auf die Schätzung der Fähigkeit bezüglich der l-ten Dimension aus, obwohl diese Dimension einen positiven Beitrag zur Bearbeitung der Items leistet!

Von zentraler Bedeutung ist die Frage, wann sich diese Paradoxie zeigt und wie häufig sie auftritt. Erstaunlicherweise lässt sich beweisen, dass der paradoxe Effekt bei jedem Test auftritt, der nicht exakte Einfachstruktur besitzt. Genauer bedeutet dies: Existiert ein Item, das Querladungen aufweist, so lässt sich zu einem gegebenen Antwortmuster $\boldsymbol{x_2}$ stets ein „schlechteres“ Antwortmuster $\boldsymbol{x_1}$ ($\boldsymbol{x_2} > \boldsymbol{x_1}$) finden, das bezüglich (mindestens) einer latenten Dimension einen höheren Schätzwert liefert. Wenn wichtige Entscheidungen basierend auf dem Schätzwert dieser latenten Dimension getroffen werden, dann wäre mitunter ein schlechteres Abschneiden ($\boldsymbol{x_1}$ an Stelle von $\boldsymbol{x_2}$) für die zu diagnostizierende Person von Vorteil gewesen.

[9]Für Ladungen von Null muss die vorangegangene Bemerkung entsprechend abgeschwächt werden.

5.2 Paradoxien des modellbasierten Schätzers in einem realen Datensatz

Wir illustrieren den beschriebenen Effekt anhand eines realen Datensatzes. Der Fragebogen zum Screening psychischer Störungen im Jugendalter (SPS-J) erfasst anhand von $k = 32$ Items vier Dimensionen (Hampel und Petermann, 2006). Diese vier Dimensionen (Subtests) messen Ängstlichkeit/Depressivität (Faktor 1), aggressives-dissoziales Verhalten (Faktor 2), Selbstwertprobleme (Faktor 3) sowie Ärgerkontrollprobleme (Faktor 4). Die (32×4) Ladungsmatrix ist nichtnegativ[10], d.h. jede Dimension steht in einem positiven (genauer: nichtnegativen) Zusammenhang zu dem jeweils erzielten Itemscore. Aus Platzgründen geben wir die exakte Ladungsmatrix nicht wieder und verweisen an dieser Stelle auf Tabelle 3 in Hampel und Petermann (2006). Ferner sei darauf hingewiesen, dass wir mit diesem Beispiel lediglich die Unangemessenheit des modellbasierten Schätzers demonstrieren wollen. Uns ist bewusst, dass dieser Test in der Praxis nicht nach dem modellbasierten Schätzer ausgewertet wird.

Basierend auf dem Antwortmuster $\boldsymbol{x}$ einer Person sei $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)^T$ der errechnete Schätzwert für die vier Dimensionen. Das Paradoxon kann nun anhand der folgenden (zunächst fiktiven) Fragestellung festgemacht werden: Wie würde sich der Schätzer verändern, wenn die Person auf dem ersten Item einen höheren Score erzielt hätte? Da jedes Item auf den vier latenten Dimensionen nichtnegativ lädt, wäre zu erwarten, dass ein höherer Score auf dem ersten Item den Schätzwert für jede Dimension erhöht bzw. zumindest nicht verringert. Im konkreten Fall frägt das erste Item nach dem Drogen/Alkoholkonsum des Jugendlichen. Es wäre folglich angemessen, wenn eine Beantwortung dieser Frage im Sinne einer höheren Ausprägung, auch den Schätzwert für jede Dimension erhöht bzw. nicht verringert.

Dies ist aber nicht der Fall. Wie sich leicht anhand der Formel des Bartlett-Scores sowie

[10]Genau genommen besitzt eines der 32 Items eine (geringfügig) negative Ladung auf einer Dimension. Dies ändert jedoch qualitativ nichts an den folgenden Analysen.

der Ladungsmatrix ausrechnen lässt[11], erhöht ein höherer Score auf diesem Item (bei sonst gleichem Antwortmuster) zwar den Schätzer für die zweite Dimension, verringert aber zugleich die Schätzwerte für alle übrigen Dimensionen. Letzteres kann absurde Konsequenzen für eine Individualdiagnostik nach sich ziehen: Ein erhöhter Score auf dem ersten Item, d.h. „erhöhter Drogenkonsum“, kann dazu führen, dass der Schätzwert für die Dimension „Selbstwertprobleme“ (aber auch für die Dimension „Ärgerkontrollprobleme“) fällt. Mit anderen Worten: Eine Person, bei der Selbstwertprobleme aufgrund ihres Antwortmusters diagnostiziert wurden, hätte dies evtl. durch die Angabe eines höheren Drogenkonsums vermeiden können.

An dieser Stelle möchten wir ausdrücklich hervorheben, dass dieser Fall keine konstruierte Pathologie ist. Für den vorliegenden Datensatz lässt sich die Argumentation auf *alle* 32 Items ausdehnen. In diesem Datensatz führt eine Erhöhung des Score bei einem *beliebigen* Item stets zu einer Verringerung des Schätzwertes in mindestens einer Dimension, obwohl alle Dimension positiv (nichtnegativ) zur Bearbeitung der Items beitragen.

Eine offensichtliche, weitreichende Problematik ergibt sich für Cut-off-basierte Entscheidungsregeln. Wird eine Person beispielsweise als „depressiv“ klassifiziert, wenn ihr Schätzwert $\hat{\theta}_1$ eine bestimmte, vorher festgelegte Schranke c_1 überschreitet, so kann es vorkommen, dass sich eine Erhöhung des Itemscores auf einem Item, welches Depressivität und eine weitere Dimension misst, positiv auf die Klassifikation auswirkt. Eine Person wäre dann als „nicht-depressiv“ klassifiziert worden, wenn sie auf einem Item, welches auch Depression misst, einen höheren Score erzielt hätte. Die Argumentation für dieses Phänomen lässt sich wie im bereits erwähnten (fiktiven) Beispiels eines Physiktests durchführen. Ein höherer Testscore auf einem Item wirkt sich lediglich auf den dominanten Faktor positiv aus. Für die übrigen von dem Item gemessenen Faktoren können kontraintuitive Veränderungen durch Erhöhung des Scores entstehen.

[11] Wir nutzen für die folgenden Analysen lediglich den Kleinste-Quadrate-Schätzer, da die Messfehlervarianzen der Items anhand der gegebenen Daten nicht direkt ersichtlich sind. Der Kleinste-Quadrate-Schätzer ist erwartungstreu und stimmt zudem bei über die Items gering variierenden Messfehlervarianzen mit dem Bartlett-Score näherungsweise überein.

Abschließend soll noch die Stärke des paradoxen Effekts kurz thematisiert werden. Zur Kennzeichnung der Stärke betrachten wir für jedes Item den maximalen Zuwachs der durch eine Erhöhung des Itemscores des entsprechenden Items um eine Einheit erzielt werden kann sowie den maximalen paradoxen Effekt (die maximale Verringerung), der durch diese Erhöhung induziert wird, d.h. wir vergleichen den Zuwachs in der Dimension, die am meisten „profitiert“, mit der Abnahme in der Dimension, die am stärksten reduziert wird. Um eine einfache Kenngröße zu erhalten, bilden wir für jedes Item den Quotienten aus maximal induzierter Abnahme (Zähler) und maximal induzierter Zunahme (Nenner). Für den vorliegenden Datensatz liegt der kleinste Quotient - korrespondierend zu dem Item mit dem geringsten (relativen) paradoxen Effekt - bei 8% und der größte Quotient - korrespondierend zu dem Item mit dem stärksten (relativen) paradoxen Effekt - bei 85%. Median und Mittelwert der Quotienten befinden sich bei 34% bzw. bei 38%. Eine Person wird somit durch eine (um eine Einheit) bessere Antwort bei einem Item in einer Dimension θ_l „belohnt“, erfährt aber gleichzeitig eine Abnahme in einer anderen Dimension θ_j $(j \neq l)$, die im Regelfall etwa ein Drittel von der Zunahme in der Dimension θ_l beträgt. Dies sollte einen groben Eindruck über die Stärke des (relativen) paradoxen Effekts in diesem Datensatz vermitteln.

Bevor wir im nächsten Abschnitt (mögliche) Konsequenzen dieses Effekts für die Praxis der Testkonstruktion erläutern, soll abschließend noch einmal das zentrale Resultat bezüglich des modellbasierten Schätzers im (mehrdimensionalen) Modell der Faktorenanalyse wiedergegeben werden: Ausgehend von einem beliebig hochdimensionalen (mindestens zweidimensionalen) faktorenanalytischen Modell mit nichtnegativer Ladungsmatrix - d.h. jede Dimension trägt zur Beantwortung eines Items positiv (präziser: nichtnegativ) bei - lässt sich das starke Resultat herleiten, dass der paradoxe Effekt jedem mehrdimensionalem Test insofern dieser nicht eine Einfachstruktur aufweist, inhärent ist. Folglich gibt es für eine *beliebige* Ladungsmatrix mit nichtnegativen Einträgen in der mindestens ein Item auf mehreren Dimensionen lädt, stets ein Item, bei dem sich eine Erhöhung des Scores unter Konstanthaltung der anderen Itemscores negativ auf die Fähigkeitsschätzung einer Dimension auswirkt

(Theorem 5, Jordan und Spiess, 2011).

Ein Test mit Einfachstruktur (der Paradoxie-freie Fall) repräsentiert zudem keinen „echt“ mehrdimensionalen Test, da dessen Items in mehrere eindimensionale Skalen aufteilbar wären. Somit impliziert unter den obigen Annahmen (faktorenanalytisches Modell mit nicht-negativer Ladungsmatrix) jeder echt mehrdimensionale Test einen paradoxen Effekt.

Um die volle Allgemeinheit des Paradoxons herauszustellen, sei ebenso bemerkt, dass ein analoges Resultat für probabilistische Modelle gilt. So weist z.B. auch jeder M2PL- (multidimensional two parameter logistic model) Test den paradoxen Effekt auf, insofern dessen Ladungsmatrix nicht Einfachstruktur besitzt.

Zudem sei erwähnt, dass ähnliche Paradoxien im Bereich des Hypothesentestens bzw. für die Intervallschätzung auftreten. Die Nullhypothese „$\theta_1 \leq \kappa$“ kann dann z.B. verworfen werden, wenn eine richtige Antwort einer Person in eine falsche Antwort umgewandelt wird.

Ein extremes Beispiel für diesen Sachverhalt ist in einem zweidimensionalen Rasch-Modell konstruierbar, in dem beide Fähigkeiten positiv zur Beantwortung der Fragen beitragen. Hier ist es bezüglich obiger H_0 möglich, durch Veränderung *einer* richtigen Antwort einen extrem hohen p-Wert (beliebig nahe an 1) auf einen p-Wert nahe Null zu reduzieren[12]. Dies erscheint besonders bemerkenswert, da der Hypothesentest in diesem mehrdimensionalen Rasch-Modell nach dem Neymann-Pearson-Lemma konstruiert und somit aus statistischer Perspektive „optimal“ ist (siehe z.B. Theorem 4.4.1 aus Lehmann und Romano, 2005).

6 Schlussfolgerungen und Konsequenzen für die Diagnostik

Auch wenn der paradoxe Effekte gravierende Folgen im Hinblick auf eine faire Individualdiagnostik aufweist, so könnte man dennoch seine Relevanz für die Praxis bezweifeln, da die

[12]Das Beispiel kann von den Autoren angefordert werden.

Auswertung in der Praxis überwiegend nicht mit dem modellbasierten, paradoxen Schätzer vorgenommen wird. Diese scheinbare praktische Irrelevanz ist jedoch ein Trugschluss. Bei Gültigkeit des (mehrdimensionalen) faktorenanalytischen Modells ist der Summenwert - im Gegensatz zum paradoxen, modellbasierten Schätzer - i.A. kein adäquater Schätzer (verzerrt, nicht suffizient, inkonsistent) für die unbeobachtbaren Ausprägungen der latenten Konstrukte[13]. Ein potentieller Einwand für diese Argumentation könnte sich darauf berufen, dass für diagnostische Zwecke in einigen Fällen lediglich die Rangordnung der Personen bezüglich der latenten Konstrukte wiedergegeben werden muss. Wenn also der einfache Summenwert hoch mit dem modellbasierten Schätzer korreliert, dann bildet er für die Ordnung der Individuen einen geeigneten und fairen Schätzwert. Letzteres sollte jedoch aus mehreren Gründen relativiert werden: Zum Einen fällt die Korrelation des Summenwertes mit dem modellbasierten Schätzer je nach Gestalt der Ladungsmatrix verschieden aus. Für den beschriebenen Datensatz liegen die Korrelationen[14] der Summenwerte mit ihren modellbasierten Pendants zwischen $\rho = 0.65$ (Dimension 4) und $\rho = 0.99$ (Dimension 3). Während sich folglich der modellbasierte Schätzer für die dritte Dimension nahezu perfekt linear anhand des einfachen Summenwerts vorhersagen lässt, erklärt der einfache Summenwert der vierten Subskala nicht einmal die Hälfte ($0.65^2 \approx 0.42$) der entsprechenden Variation des modellbasierten Schätzers. Mit anderen Worten: Die selten explizit ausgedrückte Annahme, dass der Subskalen-spezifische Summenwert eine adäquate Wiedergabe der Rangordnung der Individuen ermöglicht, sollte anhand der konkret vorliegenden Ladungsmatrix überprüft werden. Den Autoren ist jedoch nicht bekannt, dass diese (leicht durchführbare) Überprüfung in der Praxis vorgenommen wird. Aber auch wenn diese Überprüfung positiv ausfällt (d.h. hohe Korrelationen der Subskalen-spezifischen Summenwerte mit ihren modellbasierten Pendants), so sind nur „univariate“ Rangordnungen davon betroffen. Betrachtet man hingegen

[13]Die Schätzung im faktorenanalytischen Modell verläuft vollkommen analog zur Schätzung in einem (multiplen) linearen Regressionsmodell. Insofern ließe sich die Frage aufwerfen, warum Anwender in einem Modell den Summenwert als Schätzer wählen und zugleich in einem äquivalenten(!) Modell den (gewichteten) Kleinste-Quadrate-Schätzer verwenden.

[14]Zur Berechnung wurden gleiche Messfehlervarianzen der Items unterstellt.

die Ordnung der Personen im kompletten (für das Beispiel des SPS-J vierdimensionalen) Merkmalsraum, so sieht man unmittelbar anhand der charakteristischen Eigenschaft des paradoxen Effekts, dass das Prinzip der Ordnung im *mehrdimensionalen* Raum sich drastisch unterscheidet: Person A, die in einer Aufgabe einen höheren Score erzielt als Person B wird (bei sonst gleicher Performance) vom Summenscore im Merkmalsraum „höher“ angeordnet - d.h. als insgesamt fähiger betrachtet - als Person B, während der modellbasierte Schätzer keine Person als „fähiger“ klassifiziert. Betrachtet man für diese Personen folglich den gesamten vierdimensionalen Schätzer - statt Korrelationen, die lediglich univariat auf eine Dimension gerichtet sind - so gibt es stets mindestens eine Dimension in der die übliche Anordnung gemäß Summenwerten zu „falschen“ Ergebnissen führt - eine wichtige Beobachtung, die von der Berechnung der Interkorrelationen verdeckt wird. Wenn wir zwei Personen als *vergleichbar* definieren, falls eine Person in jedem latenten Merkmal mindestens so hohe (geschätzte) Ausprägungen aufweist wie die andere Person, so lässt sich die Diskrepanz zwischen modellbasiertem Schätzer und Summenscore verkürzt wie folgt ausdrücken: Personen, die sich bezüglich ihrer Testleistung (damit meinen wir die Itemscores) in eine Rangordnung bringen lassen, sind bezüglich des Summenwerts stets vergleichbar, bezogen auf den besten modellbasierten Schätzer jedoch nicht notwendigerweise. Das Modell „lehrt“ uns gewissermaßen, dass von einer umgangssprachlich „besseren“ Testleistung nicht auf eine höhere Ausprägung der zugrundeliegenden Konstrukte geschlossen werden kann. Es sei ferner darauf hingewiesen, dass die Gültigkeit des Modells im Konstruktionsprozess des Tests, d.h. bei der Auswahl und Zuordnung der Items, unterstellt wird. Aber genau jenes Modell „diktiert“ auch eine sinnvolle Schätzmethode (genauer gesagt: die „beste“ Art, die Personenparameter zu schätzen, wenn das Modell gilt). Zweifel an der Angemessenheit des modellbasierten Schätzers (im SPS-J: Ist es akzeptabel bei höherem Drogenkonsum auf geringere Selbstwertprobleme zu schließen?) hinterfragen somit implizit immer auch das zur Testkonstruktion zugrundelegte Modell.

Anstatt nach „technischen Ausflüchten“, wie z.B. die Nutzung des Summenwerts mit Be-

rufung auf eine (potentiell) approximative Einfachstruktur der Datenmatrix, zu suchen, sollte man dieser Eigenschaft des Modells Rechnung tragen. So könnte beispielsweise an Stelle einer routinierten Nutzung von Modellen, deren Aufkommen größtenteils mit ihrer mathematischen Einfachheit zu erklären ist, eine axiomatische Begründung von (neuen) testtheoretischen Modellen - ähnlich wie im Falle des Rasch-Modells - erfolgen. Diese Herangehensweise, d.h. zuerst die Auflistung von sinnvollen Eigenschaften, die ein testtheoretisches Modell aufweisen sollte (z.B. Fairness der Diagnose) und danach die Herleitung der Modellklasse, erscheint uns wesentlich produktiver (aber auch schwerer) als die algorithmische Anwendung von historisch etablierten Verfahren zur Skalenkonstruktion.

Darüberhinaus ließe sich auch der Standpunkt vertreten, dass das Paradoxon (lediglich) ein weiteres Mosaik im Hinblick auf Problematiken darstellt, die Modellen mit latenten Variablen inhärent sind. Der paradoxe Effekt wäre dann neben dem Problem der Beliebigkeit in der Interpretation der Faktoren (Rotationsproblem) und dem Problem der Vermengung von inter- und intraindividuellen Ebenen („stochastic subject view“ vs. „random sampling view“, siehe z.B. Holland, 1990) ein weiterer Punkt in der Kritik von Modellen mit latenten Variablen.

Literaturverzeichnis

Finkelman, M., Hooker, G. und Wang, Z. (2010). Prevalence and magnitude of paradoxical results in multidimensional item response theory. Journal of Educational and Behavioral Statistics, 35, 744-761.

Fischer, G. H. (1995). Derivations of the Rasch model. In Fischer G. H. und Molenaar, I. W., Rasch models: Foundations, recent developments, and applications, 15-38. New-York: Springer.

Hampel, P. und Petermann, F. (2006). Fragebogen zum Screening psychischer Störungen

im Jugendalter (SPS-J). Zeitschrift für Klinische Psychologie und Psychotherapie, 35, 204-214.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. Psychometrika, 55, 577-601.

Hooker, G., Finkelman, M., und Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. Psychometrika, 74, 419-442.

Jordan, P. und Spiess, M. (2011). Generalizations of paradoxical results in multidimensional item response theory. Psychometrika. Im Druck.

Lehmann, E. L. und Casella, G. (1998). Theory of point estimation (2nd ed.). New York: Springer.

Lehmann, E. L. und Romano, J. P. (2005). Testing statistical hypotheses (3rd ed.). New York: Springer.

Mardia, K. V., Kent, J. T. und Bibby, J. M. (1979). Multivariate Analysis. London: Academic Press.

McNemar, Q. (1946). Opinion-attitude methodology. Psychological Bulletin, 43, 289-374.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In Fischer, G. H. und Molenaar, I. W., Rasch models: Foundations, recent developments, and applications, 3-14. New-York: Springer. Rasch models - Foundations, recent developments, and applications. New York: Springer.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 49, 425-435.

Salgueiro, M. F., Smith, P.W.F. und McDonald, J.W. (2008). The manifest associa- tion structure of the single-factor model: insights from partial correlations. Psychometrika, 73, 665-670.

Shapiro, A. (1982). Rank reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. Psychometrika, 47, 187-199.

Ten Berge, J. M. F. und Socan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. Psychometrika, 69, 613-625.

1 de Jong, R., van Buuren, S. & Spiess, M. (2013). *Multiple imputation of predictor variables using generalized additive models.*

2 Jordan, P. & Spiess, M. (2013). *Fundamentale Probleme beim Einsatz testtheoretischer Modelle zur Diagnose von Individuen.*