Aus dem Psychologischen Institut
der Ruprecht-Karls-Universität Heidelberg

Multivariate Methoden der Testkonstruktion

Dissertation

zur Erlangung des Doktorgrades der
Fakultät für Verhaltens- und Empirische Kulturwissenschaften
der Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Safir Yousfi

geboren in Heidelberg
am 25.11.1969

Heidelberg, im Juni 2003

Gutachter:

Prof. Dr. Manfred Amelang
Prof. Dr. Joachim Werner

Danksagung

All denen, die zum Gelingen der vorliegenden Arbeit beigetragen haben, möchte ich ganz herzlich danken.

An erster Stelle gilt mein Dank Herrn Prof. Dr. Manfred Amelang für die Betreuung der Arbeit und die Überlassung von Daten für die empirischen Auswertungen.

Herrn Prof. Dr. Joachim Werner danke ich für die zahlreichen, nützlichen Korrekturvorschläge bei der Finalisierung und Straffung der endgültigen Version der Dissertation.

Nicht zuletzt möchte ich mich bei meiner Familie für die liebevolle und tatkräftige Unterstützung bedanken.

Vorwort

Die vorliegende Arbeit widmet sich der Frage, wie sich Methoden zur Auswahl von Items bei der Testkonstruktion auf die Testgüte auswirken. Insbesondere sollen solche Methoden der Itemselektion untersucht werden, die nicht auf dem Vergleich von Itemkennwerten beruhen, sondern die Güte verschiedener Itemkombinationen vergleichen. Diese Methoden berücksichtigen, dass sich der Nutzen eines Items für die Sicherung der Testgüte nicht unabhängig davon beurteilen lässt, welche Items sonst im Test enthalten sind. Es kommt also nicht nur darauf an, wie gut die einzelnen Items sind, sondern auch darauf, wie gut die Items zueinander passen. Da die Anzahl möglicher Itemkombinationen exponentiell mit dem Umfang des Itempools ansteigt, lassen sich jedoch selbst bei der heutigen Rechnerkapazität in der Regel nicht alle denkbaren Itemkombinationen direkt untersuchen. Es gibt jedoch Algorithmen, die gezielt nach solchen Itemkombinationen suchen, bei denen eine besonders hohe Validität zu erwarten ist. Dass dabei unter Umständen auch solche Items ausgewählt werden, deren Itemkennwerte keine besonders günstigen Werte erreichen, lässt sich durch Suppressionseffekte erklären.

Suppressionseffekte werden bei herkömmlichen Methoden der Testkonstruktion in der Regel nicht berücksichtigt. Die üblichen Itemanalysen zielen vorwiegend darauf ab, die Homogenität eines Tests zu sichern. Welcher der beiden Ansätze zu Tests höherer Reliabilität und Validität führt, dürfte nicht zuletzt von den statistischen Eigenschaften des Itempools abhängen. Der Vergleich der verschiedenen Methoden der Itemselektion anhand empirischer und simulierter Datensätze bilden den Schwerpunkt der vorliegenden Arbeit. Dabei steht die Konstruktion von Skalen aus ungewichteten Items anhand klassischer Testmodelle im Zentrum der Aufmerksamkeit. Im einzelnen gliedert sich die Arbeit wie folgt:

Zunächst sollen die Grundlagen der klassischen Testtheorie in einem eigenen Kapitel dargestellt werden. Kapitel 1.1 und 1.2 dienen vorwiegend dazu, Begriffe zu definieren und Konzepte zu erläutern, auf die in den folgenden Kapiteln zurückgegriffen wird.

In den darauf folgenden Abschnitten sollen Methoden der Itemselektion diskutiert werden. Dabei werden insbesondere auch Methoden diskutiert, die sich nicht auf die Selektion anhand von Itemkennwerten beschränken, sondern versuchen, die optimale Auswahl von Items dadurch sicherzustellen, dass auch die Korrelationen zwischen den Items berücksichtigt werden. Dabei wird auf Ansätze zur Selektion von Prädiktoren zurückgegriffen, wie sie in der multiplen Regression üblich sind. Diese multivariaten Methoden der Testkonstruktion knüpfen an Cattells

Ansatz einer strukturellen Psychometrie an. Vorteile gegenüber der herkömmlichen Methode der Itemkonstruktion resultieren dabei vor allem aus der Nutzung von Suppressoreffekten.

In einem eigenem Kapitel wird eine mathematische Präzisierung des bisher nur sehr unscharf definierten Suppressorbegriffs in der Testkonstruktion vorgeschlagen.

Im empirischen Teil der Arbeit sollen die herkömmlichen Methoden der Testkonstruktion anhand empirischer sowie simulierter Datensätze mit multivariaten Methoden verglichen werden. Die Verwendung von Monte-Carlo Simulationen ermöglicht es, gezielt zu überprüfen, unter welchen Bedingungen die verschiedenen Methoden der Itemkonstruktion indiziert sind. Die externe Validität dieser Analysen kann freilich nur anhand empirischer Datensätze beurteilt werden.

Im Diskussionsteil werden die Ergebnisse der empirischen Arbeiten unter Berücksichtigung der Ergebnisse der theoretischen Analysen interpretiert und Empfehlungen für die Testkonstruktion abgeleitet.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

1 Grundlagen der klassischen Testtheorie

In früheren Darstellungen der klassischen Testtheorie (Gulliksen, 1950) wurde versucht, fundamentale Annahmen der klassischen Testtheorie in Form von Axiomen zu formulieren. Moderne Darstellungen der klassischen Testtheorie (Steyer & Eid; 1993, Krauth, 1995) betonen, dass die meisten grundlegenden Aussagen der klassischen Testtheorie keinen empirischen Gehalt haben, sondern lediglich Definitionen sind. Mit Hilfe allgemeiner mathematisch-stochastischer Gesetzmäßigkeiten werden dann Sätze über Zusammenhänge zwischen den in den Definitionen eingeführten Begriffen abgeleitet. Diese Sätze sind tautologisch, d.h. sie können sich aus theoretischen Gründen in keiner empirischen Anwendung als falsch erweisen. Sätze, die empirisch getestet werden können, erhält man erst dann, wenn man zusätzliche Modellspezifikationen einführt, die empirischen Gehalt haben (Steyer & Eid, 1993; Rost, 1988). In Anlehnung an Steyer und Eid (1993) soll in der folgenden Darstellung der klassischen Testtheorie eine klare Trennung zwischen solchen Sätzen der klassischen Testtheorie, die tautologisch sind, und denjenigen, die empirischen Gehalt haben, eingehalten werden. Diese Unterscheidung ist von zentraler Bedeutung, da Aussagen einer Theorie, die empirischen Gehalt haben, bei der Anwendung auf einen Gegenstand nach Möglichkeit einer empirischen Prüfung zugänglich gemacht werden sollten. Bei tautologischen Aussagen sind solche Anstrengungen völlig sinnlos, da sie immer wahr sein müssen. Werden Sätze und Theoreme dagegen aus Aussagen abgeleitet, die empirischen Gehalt haben, so gelten sie nur dann zwingend, falls die zugrundeliegenden Aussagen in einer konkreten empirischen Situation tatsächlich erfüllt sind. Auf die mathematische Herleitung der einzelnen Sätze wird in diesem Kapitel aus Gründen der Übersichtlichkeit weitgehend verzichtet und auf die entsprechende Literatur verwiesen. Insbesondere sei hier auf die Darstellung von Steyer und Eid (1993) verwiesen, die von der formalen Ausarbeitung ein sehr hohes Niveau hat.

1.1 Grundlegende Begriffe

Die Vorgabe eines psychologischen Tests wird im Rahmen der klassischen Testtheorie als Zufallsexperiment aufgefasst, da die Antwort der Person – und in vielen Fällen auch die Auswahl der Person – nicht streng determiniert sein dürfte. Es wird also davon ausgegangen, dass jedes mögliche Antwortverhalten einer Person in der konkreten Testsituation mit einer gewissen Wahrscheinlichkeit (u.U. auch Null) auftreten könnte. Meist werden den Antworten der Personen bei der Auswertung eines Tests rationale Zahlen zugeordnet. Im einfachsten Fall sind der Testperson oder der Person, die den Test auswertet, nur zwei verschiedene Alternativen vorgegeben, z.B. Zustimmung oder Ablehnung zu einer Aussage, korrekte oder fehlerhafte

Bearbeitung einer Testaufgabe. Zählt man die Anzahl der zustimmenden oder korrekten Antworten einer Person und fasst jede Antwort als Realisierung eines Zufallsexperiments auf, so ist damit eine diskrete Zufallsvariable definiert, die nur natürliche Zahlen als Werte annehmen kann. Geht man davon aus, dass die Wahrscheinlichkeiten jedes möglichen Wertes dieser Zufallsvariable bekannt sind, so ließe sich der Erwartungswert dieser Zufallsvariable berechnen. Der (bedingte) Erwartungswert dieser Zufallsvariable X wird im Rahmen der klassischen Testtheorie als wahrer Wert τ der Person u bezeichnet (vgl. Lord & Novick, 1968, S. 30; Steyer & Eid, 1993, S. 109; Krauth, 1995, S. 239).

$$\tau := E\left[X \middle| p_u \right]$$

[1.1-1]

Spielen bei der Auswahl der Person(en) zufällige Einflüsse eine Rolle, so ist τ eine stochastische Variable. p_u ist eine stochastische Variable, deren Ausprägung anzeigt, welche Person getestet wurde. Die Wahrscheinlichkeit, dass eine bestimmte Person u ausgewählt wurde kann – z.B. in einer diagnostischen Situation – auch eins sein.

Diese Definition setzt implizit voraus, dass der Erwartungswert der entsprechenden Zufallsvariable existiert. Macht man die Annahme, dass das zweite Moment dieser Zufallsvariable existiert und endlich ist, so lässt sich herleiten, dass nicht nur der Erwartungswert, sondern auch die Varianz der Zufallsvariable existieren und endlich sind (Krauth, 1995). Diese implizite Annahme hat zwar empirischen Gehalt, so dass es gerechtfertigt ist, sie als Existenzaxiom zu bezeichnen (Moosbrugger, 1988). In praktischen Anwendungen ist es aufgrund der Natur der untersuchten Variablen jedoch meist angemessen, sie einfach vorauszusetzen. Für die Definition weiterer Begriffe der klassischen Testtheorie (z.B. Reliabilität) ist diese Annahme nicht nur sinnvoll, sondern notwendig. Die Definition des wahren Wertes ([1.1-1]) ist dagegen lediglich eine alternative Benennung des Begriffs der bedingten Wahrscheinlichkeit.

Die Zufallsvariable X, die das Verhalten der Person in dem obigen Beispiel beschrieben hat, kann nur natürliche Zahlen als Werte annehmen, während der Erwartungswert dieser Zufallsvariable, also der wahre Wert, im Allgemeinen keine natürliche Zahl sein dürfte. Der wahre Wert der Person ist hier also ein Wert, der in dem Zufallsexperiment selbst gar nicht vorkommen konnte. Man muss bei der anschaulichen Deutung des wahren Wertes also sehr vorsichtig sein (Krauth, 1995). Er soll zwar die Verhaltenstendenz eines Individuums in einer konkreten Situation kennzeichnen; er ist aber nur ein Kennwert einer Wahrscheinlichkeits-

funktion und sollte nicht als die "wahre", unverfälschte, von zufälligen Fehlereinflüsse unbeeinflusste Reaktion des Individuums verstanden werden (Steyer & Eid, 1993). Dieser problematischen Interpretation wird aber nicht nur durch die unglückliche Begriffswahl Vorschub geleistet, sondern auch durch die Konvention, die Differenz zwischen dem wahren Wert und dem Rohwert – damit ist der tatsächlich beobachtete Testwert gemeint – als Messfehler ε zu bezeichnen (vgl. Lord & Novick, 1968, S. 31; Steyer & Eid, 1993, S. 109; Krauth, 1995, S. 239).

$$\varepsilon := X - \tau$$

[1.1-2]

Moosbrugger (1988) bezeichnet die Tatsache, dass der Rohwert sich additiv aus wahrem Wert und Messfehler zusammensetzt, als Verknüpfungsaxiom. Aus den Definitionen von wahrem Wert ([1.1-1] auf S. 2) und Fehler ([1.1-2]) sowie dem Existenzaxiom lässt sich jedoch leicht ableiten, dass der Rohwert die Summe von wahrem Wert und Messfehler sein muss (Krauth, 1995).

Will man den Messfehler nicht nur formal definieren, sondern inhaltlich interpretieren, etwa als die "Gesamtheit aller unsystematischen und nicht kontrollierbaren oder vorhersagbaren Einflussgrößen, die auf das Messergebnis einwirken können" (Amelang & Zielinski, 1997, S. 34), so ist zu beachten, dass diese Einflussgrößen so beschaffen sein müssen, dass sie prinzipiell nicht der Gegenstand eines anderen Tests sein könnten. Werden beispielsweise die Ergebnisse eines Tests zur Erfassung der Intelligenz vom *aktuellen* Konzentrationsniveau beeinflusst, so ist es nicht zulässig, zu behaupten, der wahre Wert des Tests sei die Intelligenz der Testperson, während das *aktuelle* Konzentrationsniveau Teil des Messfehlers ist. Würde man nämlich das *aktuelle* Konzentrationsniveau mit Hilfe eines anderen (State-)Tests erfassen, so wären die wahren Werte dieses Tests mit den Fehlerwerten des Intelligenztests korreliert. Es lässt sich jedoch allein aus den oben gegebenen Definitionen und allgemeinen stochastischen Gesetzen herleiten, dass diese Aussage logisch (!) falsch ist[1] (siehe [1.1-9] auf S. 6 sowie Krauth, 1995, S. 242; Steyer & Eid, 1993, S. 110). In den wahren Werten des Intelligenztests drückt sich

[1] Der logische Widerspruch ergibt sich jedoch nur dann, wenn man bei der Definition der wahren Werte beider Tests fordert, dass beide Werte bedingte Wahrscheinlichkeiten bezüglich derselben Bedingung sind. Definiert man den wahren Wert des Intelligenztests als den Erwartungswert des Intelligenztests über alle Zeitpunkte und Situationen hinweg, während der wahre Wert des Konzentrationstests nur der Erwartungswert zu einem bestimmten Zeitpunkt in einer spezifischen Situation sein soll, so wäre es in der Tat denkbar, dass der wahre Wert des einen Tests mit dem Fehler des anderen korreliert. In diesem Fall verlieren jedoch alle Aussagen, die in der klassischen Testtheorie über die Zusammenhänge zweier Tests gemacht werden, ihre Gültigkeit (u.a. auch [1.1-9] auf S. 6).

demnach nicht nur die Intelligenz aus – als *trait* –, sondern auch die Konzentration – als *state* –. Diese Überlegung schränkt eine inhaltliche Interpretation des Messfehlers stark ein, da nahezu jede Einflussgröße, die in nachvollziehbarer Weise Einfluss auf das Messergebnis hat, selbst Gegenstand einer Messung sein könnte.

Eine angemessenere Interpretation des Messfehlers ergibt sich vielleicht, wenn man davon ausgeht, dass das Verhalten der Messobjekte seiner Natur nach stochastisch ist. Damit ist gemeint, dass selbst bei vollständiger Kenntnis sämtlicher Einflussgrößen das Verhalten der Person nicht genau vorhergesagt werden kann, sondern bestenfalls Wahrscheinlichkeiten einzelner Verhaltensweisen präzise angegeben werden können. Der Messfehler wäre demnach weniger ein Ausdruck der Unzulänglichkeit des Messinstrumentes als vielmehr die Konsequenz der natürlichen Variabilität des Verhaltens, und damit kein Fehler im eigentlichen Sinn. Die stochastische Konzeptualisierung des Forschungsgegenstands ist selbst in der Physik – dem Paradebeispiel einer exakten Wissenschaft – seit längerer Zeit in weiten Teilen üblich. Streng deterministische Vorstellungen, wie sie etwa noch von Laplace (1814) formuliert wurden, sind in der Physik im Zuge der Entwicklung der Quantenmechanik durch stochastische Modelle abgelöst worden (Heisenberg & Bohr, 1963). In der Psychologie wird der Forschungsgegenstand zwar auch meist mit stochastischen Theorien modelliert und Daten statistisch ausgewertet. Meist lastet man den damit verbundenen Mangel an Präzision jedoch weniger dem Gegenstand selbst an, sondern führt ihn – ganz in der Tradition von Laplace (1814) – auf unzureichende Information oder die Unzulänglichkeit der eigenen Methoden zurück.

Weitere Folgerungen, die sich allein aus der Existenz von Erwartungswert und Varianz des Testwertes ergeben sind (vgl. Krauth, 1995; Steyer & Eid, 1993):

$$E\left[\varepsilon \middle| p_u\right] = 0 \qquad \textbf{[1.1-3]}$$

$$E\left(\varepsilon\right) = 0 \qquad \textbf{[1.1-4]}$$

$$\sigma_{\tau,\varepsilon} = 0 \qquad \textbf{[1.1-5]}$$

$$\sigma^2\left[\varepsilon \middle| p_u\right] = \sigma^2\left[X \middle| p_u\right] \qquad \textbf{[1.1-6]}$$

$$E(X) = E(\tau)$$

[1.1-7]

$$\sigma^2(X) = \sigma^2(\tau) + \sigma^2(\varepsilon)$$

[1.1-8]

Interpretiert man den wahren Wert einer Person im Sinne der Persönlichkeitspsychologie als den Grad der Ausprägung einer zugrundeliegenden Persönlichkeitseigenschaft, so handelt es sich dabei um eine operationale Definition der zugrundeliegenden Eigenschaft, da jeder Person der Erwartungswert in diesem Test als Ausprägung in der betreffenden Eigenschaft zugeordnet wird. Dabei ist jedoch zu beachten, dass sich diese Aussage nur auf das Verhalten in der spezifischen Testsituation bezieht. Die Eigenschaft wird also durch die Prozedur der Messung definiert und nicht unter Bezug auf theoretische oder inhaltliche Erwägungen. Die theoretischen und inhaltlichen Erwägungen dürften allerdings bei der Formulierung und Auswahl der Items von entscheidender Bedeutung sein. Es bleibt jedoch zunächst völlig offen, inwieweit es gelingt, durch die Festlegung der Bedingungen für das durch den Test realisierte Zufallsexperiment die intendierte Persönlichkeitseigenschaft zu messen. Weiterhin ist vorstellbar, dass bei einer Testwiederholung oder wenn man den Test unter anderen Bedingungen vorgegeben hätte, sich für die einzelnen Reaktionen der Person andere Wahrscheinlichkeiten und damit auch andere wahre Werte ergeben würden. Die Wiederholung eines Tests stellt in diesem Sinne ein neues Zufallsexperiment und damit auch einen anderen Test dar, dessen Ergebnisse völlig anders sein können. Überspitzt könnte man die operationalen Definitionen zugrundeliegende persönlichkeitspsychologische Modellvorstellung etwa so formulieren: Die Person *u* hat die Tendenz, sich in der durch Vorgabe des Tests definierten Situation *X* zum Zeitpunkt *t* den Test auf eine gewisse Art zu bearbeiten. Verallgemeinerungen auf andere Situationen und Zeitpunkte sind auf der Grundlage der bisherigen Definition des wahren Wertes also durch nichts zu rechtfertigen.

In den folgenden Abschnitten zu Modellspezifikationen und Gütekriterien psychologischer Tests wird auf Ansätze eingegangen, die eine sinnvolle weitergehende Interpretation des wahren Wertes zum Ziel haben. Hierbei spielen vor allem die Stabilität (→ Reliabilität) der Testwerte und deren statistische Zusammenhänge mit anderen Zufallsvariablen (→ externe Validität, interne Konsistenz) eine große Rolle.

1.1.1 Modellspezifikationen mit empirischem Gehalt

Wie bereits erwähnt, sind die grundlegenden Aussagen der Testtheorie weitgehend ohne empirischen Gehalt. Betrachtet man die Definitionen von wahrem Wert und Fehler, so lässt sich unschwer erkennen, dass jede Zufallsvariable, deren Erwartungswert definiert ist, einen Test im Sinne der klassischen Testtheorie darstellt. Im Bereich der Persönlichkeitspsychologie soll der wahre Wert eines Tests die Verhaltenstendenz einer Person kennzeichnen. Damit ein Test diesem Anspruch nicht nur in Bezug auf eine konkrete Situation – nämlich dem Verhalten der Person während der Bearbeitung des Tests – gerecht wird, müssen weitere Annahmen eingeführt und empirisch geprüft werden. Diese Annahmen beziehen sich auf die statistischen Zusammenhänge mit anderen Tests, also dem Verhalten – oder besser, den Verhaltenstendenzen – der Person in anderen Situationen. Wahrscheinlichkeitstheoretisch betrachtet, stellt selbst die wiederholte Vorgabe derselben Testitems eine andere Situation bzw. ein neues Zufallsexperiment und damit auch einen neuen Test dar, da sich die Verhaltenstendenzen des Individuums im Vergleich zur ersten Durchführung des Tests aus verschiedenen Gründen geändert haben können. Die Annahme, dass sich bei Wiederholung derselben Testitems dieselbe latente Eigenschaft ausdrückt, stellt bereits eine empirisch testbare Modellspezifikation dar. Gegenstand der Modellspezifikationen sind aber meist die Zusammenhänge mit den Ergebnissen von anderen Tests. Wie bereits erwähnt, kann man (fast) jede Zufallsvariable, also auch einzelne Items, als Test bezeichnen. Modellspezifikationen lassen sich also auch zur Beschreibung der stochastischen Zusammenhänge zwischen den Items eines Tests verwenden.

Bereits ohne weitere Annahmen lässt sich aus den oben formulierten Definitionen ([1.1-1] auf S. 2 und [1.1-2] auf S. 3) ableiten, dass der wahre Wert eines Tests oder Items mit dem Fehlerwert eines anderen Tests oder Items nicht korrelieren kann (vgl. [1.1-3] und [1.1-4] auf S. 4 oder auch Steyer & Eid, 1993, S. 110; Krauth, 1995, S.242):

$$\begin{aligned}\sigma(\tau_1,\varepsilon_2) &= E(\tau_1\cdot\varepsilon_2) - E(\tau_1)E(\varepsilon_2)\\ &= E(\tau_1\cdot\varepsilon_2)\\ &= E\left(E\left[\tau_1\cdot\varepsilon_2\,|\,p_u\right]\right)\\ &= E\left(\tau_1\cdot E\left[\varepsilon_2\,|\,p_u\right]\right)\\ &= E(\tau_1\cdot 0)\\ &= 0\end{aligned}$$

[1.1-9]

Hieraus folgt, dass die Kovarianz der wahren Werte gleich der Kovarianz zwischen dem wahren Wert des einen und dem beobachtetem Wert des anderen Tests ist (vgl. [1.1-9] oder auch Zimmerman & Williams, 1980):

$$\begin{aligned}\sigma(\tau_1,\tau_2) &= \sigma(X_1,\tau_2) - \sigma(\varepsilon_1,\tau_2)\\ &= \sigma(X_1,\tau_2)\\ &= \sigma(X_1,X_2) - \sigma(\varepsilon_1,\varepsilon_2)\end{aligned}$$

[1.1-10]

[1.1-9] und [1.1-10] sind ebenso wie [1.1-3] bis [1.1-8] (siehe S. 4) direkte Implikationen aus der Definition des wahren Wertes [1.1-1] (S. 2) und des Fehlers [1.1-2] (S. 3). Keine dieser Aussagen kann also in empirischen Anwendungen falsch sein. Sie haben demnach keinen empirischen Gehalt, da sie Tautologien – und damit logisch notwendigerweise wahr – sind (Westermann, 2000). Einzige Voraussetzung für die Gültigkeit von [1.1-3] bis [1.1-10] ist, dass die Begriffe, auf die sich die Aussagen beziehen, in einer konkreten empirischen Situation überhaupt definiert sind. Dies ist immer dann der Fall, wenn das erste und zweite Moment der Verteilung der entsprechenden Zufallsvariablen existieren (vgl. die Erläuterungen zu [1.1-1] auf S. 2).

Lokale Unkorreliertheit

In vielen Anwendungen der klassischen Testtheorie wird darüber hinaus gefordert, dass die Fehlerwerte einer Person bei verschiedenen Tests unkorreliert sind (Lokale Unkorreliertheit der Messfehler; Lord & Novick, 1968, S. 45; Krauth, 1995, S. 240).

$$\rho\left[\varepsilon_1,\varepsilon_2 \middle| p_u\right] = 0$$

[1.1-11]

Diese Forderung ist äquivalent zu der Aussage, dass auch die Testwerte einer Person in verschiedenen Tests unkorreliert sind (Lokale Unkorreliertheit der Testwerte, Krauth, 1995, S. 240).

$$\rho\left[X_1,X_2 \middle| p_u\right] = 0$$

[1.1-12]

Obwohl die Annahme der lokalen Unkorreliertheit der Fehler durchaus empirischen Gehalt hat, gibt es keine Methode, sie direkt zu überprüfen. Krauth (1995) empfiehlt daher, schon bei der Formulierung von Items eines Tests darauf zu achten, dass die Iteminhalte semantisch unabhängig sind, da ansonsten die Gefahr bestünde, dass die Personen (z.B. in dem Bestreben,

ein konsistentes Bild abzugeben) Antworten generieren, die nicht lokal unkorreliert sind. Aus dem bisher Gesagten sollte allerdings nicht der Schluss gezogen werden, dass es keinerlei Zusammenhänge zwischen der Bearbeitung einzelner Items geben darf. Wenn die Antwort auf ein Testitem davon abhängt, welche Items zuvor bearbeitet wurden (Reihenfolge- und Trainingseffekte), so muss dies nicht im Widerspruch zur Forderung der lokalen Unkorreliertheit stehen. Wenn beispielweise bei der Bearbeitung eines homogenen Leistungstests die letzten Items aufgrund von Lerneffekten häufiger gelöst werden, so stellt dies keine Verletzung der Annahme der lokalen Unkorreliertheit dar, sofern die Lerneffekte unabhängig davon auftreten, ob die vorangegangenen Aufgaben gelöst wurden oder nicht. Wenn allerdings allen Aufgaben dasselbe Prinzip zugrunde liegt, so dass, nachdem es einmal erkannt worden ist, die Bearbeitung der folgenden Aufgaben erleichtert wird, so wäre weder die lokale Unkorreliertheit der Messwerte noch die Unkorreliertheit der Messfehler in der Personenpopulation gegeben. Die Lerneffekte wären in diesem Fall reaktionskontingent. Modelle für diesen Fall behandelt Rost (1996).

Für die meisten Sätze der klassischen Testtheorie muss man jedoch gar nicht die sehr strenge Annahme der lokalen Unkorreliertheit machen, sondern es reicht, wenn man voraussetzt, dass die Messfehler in der betreffenden Personenpopulation nicht korrelieren (Steyer & Eid, 1993). Aus der lokalen Unkorreliertheit der Rohwerte zweier Tests folgt bei zufälliger Auswahl der Testpersonen die Unkorreliertheit der zugehörigen Messfehler in der Personenpopulation (vgl. Krauth, 1995, S. 243).

$$\rho\left[X_1, X_2 \,|\, p_u\right] = 0 \qquad \Rightarrow \qquad \rho\left(\varepsilon_1, \varepsilon_2\right) = 0$$

[1.1-13]

Aus der Unkorreliertheit der Messfehler folgt jedoch keineswegs, dass die Testwerte auch in der zugrundegelegten Personenpopulation oder in einer Stichprobe nicht korrelieren (globale Unkorreliertheit). Korrelationen zwischen den Testwerten bilden sogar die Grundlage für die Testkonstruktion im Rahmen der klassischen Testtheorie.

Für einige Sätze der klassischen Testtheorie (und einige Modelle der Faktorenanalyse) werden nicht nur Annahmen über Korrelationen der Fehler, sondern auch über die Homogenität der Fehlervarianzen gemacht. In der Regel wird dabei nur vorausgesetzt, dass die Fehlervarianz zweier Variablen in der Personenpopulation identisch ist. Außer für die Berechnung von Konfidenzintervallen in der Einzelfalldiagnostik ist es in der Regel jedoch nicht erforderlich, die strengere Annahme zu machen, dass die Fehlervarianzen verschiedener Personen identisch sind.

τ-Äquivalenz

Während die Annahme unkorrelierter Messfehler bei allen gängigen Anwendungen der klassischen Testtheorie gemacht wird, definieren die nun darzustellenden Annahmen jeweils ein spezifisches Testmodell. Diese Annahmen beziehen sich auf den Zusammenhang der wahren Werte verschiedener Tests. Fordert man, dass die wahren Werte jeder Person *u* in mehreren Tests (oder Items) dieselben sind, so ist damit das Modell τ-äquivalenter Variablen (bzw. Items) definiert (Steyer & Eid, 1993, S. 152):

$$\tau_1 := E\left[X_1 \middle| p_u\right] = E\left[X_2 \middle| p_u\right] =: \tau_2$$

[1.1-14]

Auch wenn in den Lehrbüchern zur Testtheorie selten darauf hingewiesen wird, lässt sich bei Erfüllung dieser Annahme eine Skala konstruieren, die Absolutskalenniveau hat, indem jeder Person einfach ihr wahrer Wert, also der Erwartungswert, in den zugrundeliegenden Tests zugewiesen wird. Rost (1996) erwähnt diesen Sachverhalt nur bei der Besprechung des Binomialmodells. Das Binomialmodell ist ein Spezialfall des Modells τ-äquivalenter Variablen für den Fall, dass die Rohwerte der einzelnen Items (bzw. Tests) nur zwei Werte annehmen können, wobei die Wahrscheinlichkeit einer Kategorie bei allen Items gleich ist. Diese Tatsache findet in der Literatur wohl deswegen kaum Beachtung, weil die Prozedur, mit der das Verhalten einer Person bei der Testauswertung in eine Zahl transformiert wird, in den meisten Anwendungen innerhalb der Psychologie mit einer gewissen Willkür behaftet ist (z.B. Summe der Items). Trotzdem definiert, formal betrachtet, jede dieser Prozeduren bei Geltung des Modells τ-äquivalenter Variablen eine eigene Absolutskala.

Parallelität

Fordert man, dass neben den wahren Werten auch die Fehlervarianzen für jede Person in mehreren lokal unkorrelierten Tests bzw. Items identisch sind

$$\sigma^2\left[\varepsilon_1 \middle| p_u\right] = \sigma^2\left[\varepsilon_2 \middle| p_u\right],$$

[1.1-15]

so spricht man von parallelen Tests bzw. Items (Krauth, 1995, S. 243).

Essentielle τ-Äquivalenz

Lässt man dagegen zu, dass sich die wahren Werte von je zwei Tests oder Items um einen konstanten Wert λ_{12} unterscheiden, so spricht man von dem Modell *wesentlich* (bzw. essentiell)

τ-äquivalenter Tests bzw. Items (Steyer & Eid, 1993, S. 152; Krauth, 1995, S. 248; Rost, 1996, S. 113).

$$\tau_1 = \tau_2 + \lambda_{12}$$

[1.1-16]

Verschiedene wesentlich τ-äquivalente Items unterscheiden sich im Allgemeinen hinsichtlich ihrer Schwierigkeit. Es lässt sich zeigen, dass das Modell wesentlich τ-äquivalenter Tests genau dann gültig ist, wenn eine stochastische Variable η existiert, die sich bis auf einen für den jeweiligen Test spezifischen konstanten Wert λ_t von dem entsprechenden wahren Werten unterscheidet (Steyer & Eid, 1993). Die stochastische Variable η, welche die Verhaltenstendenz der einzelnen Person in den betreffenden Testsituationen kennzeichnet, ist differenzskaliert, d.h. bis auf Translationen (Addition mit einer Konstanten) eindeutig festgelegt. Auch die Itemparameter λ_t sind differenzskaliert (Steyer & Eid, 1993).

Daraus folgt, dass Aussagen über die Varianzen der Personenparameter in der zugrundeliegenden Population bedeutsam – d.h. invariant gegenüber zulässigen Transformationen – sind. Aussagen über Differenzen sowohl zwischen den Parametern einzelner Personen ($\eta[p_u=y_1] - \eta[p_u=y_2]$) als auch zwischen den Parametern verschiedener Items ($\lambda_{t=1} - \lambda_{t=2}$) sind ebenfalls bedeutsam. Vergleiche zwischen Personenparametern hängen also ebenso wenig wie Vergleiche der Itemparameter von der gewählten Parametrisierung ab. Diese Eigenschaft nennt man spezifische Objektivität.

τ-Kongenerität

Fordert man lediglich, dass sich die wahren Werte verschiedener Tests durch positive lineare Transformationen ineinander überführen lassen, so entspricht dies dem Modell τ-kongenerischer Variablen (Steyer & Eid, 1993, S. 174 und S. 199; Rost, 1996, S. 114).

$$\tau_1 = \kappa_{12}\,\tau_2 + \lambda_{12} \qquad \text{(wobei } \kappa_{12}>0) \qquad \Leftrightarrow \qquad \rho(\tau_1, \tau_2) = 1$$

[1.1-17]

Essentielle τ-äquivalente Tests sowie Paralleltests sind immer auch τ-kongenerisch. Es lässt sich zeigen, dass Tests genau dann τ-kongenerisch sind, wenn eine stochastische Variable η existiert, so dass sich die wahren Werte aller Tests als positive, lineare Funktionen dieser Variablen darstellen lassen (Steyer & Eid, 1993, S. 199).

$$\tau_i = \kappa_i\,\eta - \lambda_i \qquad \text{(wobei } \kappa_i>0)$$

[1.1-18]

Die Personenparameter η und der Itemparameter λ_i sind intervallskaliert, d.h. sie sind eindeutig bis auf lineare Transformationen. Die Itemparameter κ_i sind verhältnisskaliert, d.h. sie sind eindeutig bis auf die Multiplikation mit einer Konstanten (Steyer & Eid, 1993). Die Itemparameter κ_i sollten nicht mit der Itemtrennschärfe (Korrelation eines Items mit dem Gesamttest) verwechselt werden, da ein Item mit einem großen Itemparameter trotzdem eine niedrige Trennschärfe haben kann, wenn die Fehlervarianz sehr groß ist.

Jede beliebige Linearkombination τ-kongenerischer Variablen definiert einen Test, für den ebenfalls τ-Kongenerität bezüglich der ursprünglichen Variablen gegeben ist. Eine Gewichtung der Items eines τ-kongenerischen Tests kann zwar zu einer Verbesserung der psychometrischen Gütekriterien führen, sie wird jedoch, entgegen den Ausführungen von Rost (1988), keineswegs durch das Modell impliziert.

Faktorenanalyse

Verallgemeinert man dass Modell τ-kongenerischer Variablen dahingehend, dass sich die wahren Werte verschiedener Tests τ_i als Linearkombination mehrerer latenter Variablen η_j (gemeinsame Faktoren) und einer für den jeweiligen Test spezifischen Einflussgröße α_i (spezifischer Faktor) darstellen lässt, so ist damit dass Modell der Faktorenanalyse definiert (Basilevsky, 1994).

$$\tau_i = \sum_j \kappa_{ij}\eta_j + \alpha_i \qquad \textbf{[1.1-19]}$$

Wenn die beobachteten Variablen X_i und die (wahren) Faktorwerte η_j z-standardisiert sind, so nennt man die Gewichtungsfaktoren κ_{ij} Faktorladungen.

Im Allgemeinen fordert man von den gemeinsamen Faktoren (η_j), zumindest aber von den spezifischen Einflussgrößen (α_i) der einzelnen Items (X_i), dass sie mit den anderen Größen des Modells unkorreliert sind. Den Varianzanteil einer Variablen, der auf die gemeinsamen Faktoren zurückgeht, bezeichnet man als Kommunalität. Der restliche Anteil der Varianz, der durch Messfehler und spezifische Faktoren verursacht wird, bezeichnet man als Spezifität. Im Gegensatz zur (essentielle) τ-Äquivalenz ([1.1-16] auf S. 10), der τ-Kongenerität ([1.1-17] auf S. 10) und der Homogenität der Fehlervarianzen ([1.1-15] auf S. 9) lässt sich das Modell der Faktorenanalyse ([1.1-19]) nur dann empirisch testen, wenn man zusätzlich Annahme über Modellparameter – z.B. über die Anzahl der latenten Faktoren – einführt (Basilevsky, 1994).

Im mehrdimensionalen Modell der Faktorenanalyse sind die wahren Werte eines Tests im Allgemeinen nicht als Funktion der wahren Werte anderer Tests darstellbar. Dies liegt an der Annahme einer spezifischen systematischen Varianzquelle für jeden Test. Der Zusammenhang zwischen den Testwerten und den latenten Variablen wird meist geschätzt, indem man die Eigenwerte und Eigenvektoren der Kovarianz- oder Korrelationsmatrix der Tests berechnet, wobei die Varianzen in der Regel durch Schätzungen der Kommunalität ersetzt werden. Dieses Verfahren ist jedoch mit einigen Unsicherheiten und willkürlichen Festlegungen behaftet (Wahl der Extraktionsmethode, Abbruchkriterium, Kommunalitätenschätzung, Rotationsmethode). Zum einen lässt sich im Allgemeinen nicht mit Sicherheit angeben, wie viele Faktoren notwendig sind, um das Verhalten der Personen in der Testsituation zu erklären. Anderseits lassen sich selbst bei einer gegebenen Anzahl von Faktoren immer noch unzählige Lösungen angeben, die hinsichtlich ihres Erklärungspotentials identisch und durch Rotationstransformationen ineinander überführbar sind. Tatsächlich definiert jede lineare Transformation der Faktoren, bei der die lineare Unabhängigkeit (oder Orthogonalität, falls die strengere Bedingung unkorrelierter Faktoren gemacht wird) erhalten bleibt, eine Menge stochastischer Variablen, die das Modell der Faktorenanalyse ebenfalls erfüllen.

Meist wird daher versucht, die Faktoren so zu rotieren, dass sich eine Einfachstruktur der Zusammenhänge zwischen Tests und Faktoren ergibt. Damit ist gemeint, das jeder Test nach Möglichkeit mit einem Faktor hoch und mit den anderen Faktoren nicht korrelieren sollte oder dass umgekehrt die Korrelationen der einzelnen Tests mit den Faktoren entweder sehr hoch oder sehr gering sein sollten. Gelingt es, die Faktoren so zu rotieren, dass jeder Test ausschließlich auf einem Faktor lädt, so erhält man ebenso viele Skalen τ-kongenerischer Variablen, wie Faktoren extrahiert wurden. In diesem (extrem seltenen vorkommenden) Fall kann man die Annahme der Multidimensionalität der Items fallen lassen. In der Praxis gestaltet sich jedoch selbst nach einer entsprechenden Rotation die Interpretation der Faktoren eher schwierig.

Die Ausprägungen der Personen auf den latenten Faktoren η_j (Faktorwerte) werden durch Linearkombinationen der Tests geschätzt. Die einzelnen Tests X_i werden also (jeweils mit dem Faktor υ_i) gewichtet und summiert, um zu adäquaten Schätzungen der Faktorwerte zu gelangen. Im Allgemeinen sind diese Schätzungen jedoch mit systematischen Fehlern behaftet, die durch die spezifischen Varianzquellen α_i der einzelnen Tests verursacht werden.

$$\begin{aligned}\hat{\eta}_{j0} &= \sum_i \upsilon_i X_i \\ &= \sum_i \upsilon_i \tau_i + \sum_i \upsilon_i \varepsilon_i \\ &= \sum_i \sum_j \upsilon_i \kappa_{ij} \eta_j + \sum_i \upsilon_i \alpha_i + \sum_i \upsilon_i \varepsilon_i\end{aligned}$$

[1.1-20]

Wenn die Schätzungen der Faktorwerte eines Faktors η_{j0} zumindest von den Werten auf den anderen gemeinsamen Faktoren unabhängig sein sollen (factor-trueness), so muss für alle $j \neq j_0$ folgende Beziehung gelten (vgl. Cattell & Tsujioka, 1964):

$$\sum_i \upsilon_i \kappa_{ij} = 0$$

[1.1-21]

Methoden, die darauf abzielen, die Faktortreue der geschätzten Faktorwerte zu sichern (z.B. Bartlett, 1937; Anderson & Rubin, 1956) maximieren im Gegensatz zur Schätzung der Faktorwerte über multiple Regression (Thurstone, 1935) nicht die Korrelation mit dem latenten Faktor (Validität). Alle hier erwähnten Methoden zur Berechnung von Faktorwerten erreichen das von ihnen angestrebte Zielkriterium (Faktortreue bzw. Validität) jedoch nur unter der Voraussetzung, dass Stichprobenfehler bei der Schätzung der Faktorladungen vernachlässigbar sind.

Verzichtet man auf eine Gewichtung (bzw. lässt man für die Gewichtungskoeffizienten nur Werte von 0 und 1 zu), so ist es selbst bei bekannten Faktorladungen im Allgemeinen nicht möglich, faktorreine Schätzungen der Faktorwerte vorzunehmen. Cattell und Tsujioka (1964) fordern jedoch, dass man zumindest versucht, den Ausdruck [1.1-21] für den jeweiligen Faktor zu minimieren. In der Praxis der Testkonstruktion wird jedoch häufig so vorgegangen, dass man zunächst eine Faktorenanalyse durchführt, bei der optimale Gewichte für die einzelnen Items geschätzt werden. Die statistischen Zusammenhänge zwischen den Items und den Faktoren bilden im Allgemeinen die Grundlage für die inhaltliche Interpretation der Faktoren. In der Endform vieler faktorbasierter Skalen wird jedoch auf eine Gewichtung der Items verzichtet. Stattdessen werden einfach die Items mit den höchsten Ladungen aufsummiert. Dies führt im Allgemeinen dazu, dass nicht mehr mit Sicherheit behaupten werden kann, dass die Skala ausschließlich den intendierten Faktor misst. Dies wäre nur dann der Fall, wenn es tatsächlich gelungen ist, eine perfekte Einfachstruktur mit mehreren τ-kongenerischen Skalen zu konstruieren, oder wenn sich die Einflüsse anderer Faktoren zufällig ausgleichen.

Auf der anderen Seite ist es unpraktikabel und in vielen Anwendungen unerwünscht, wenn man die optimal gewichtete Summe aller Items eines mehrdimensionalen Tests bilden muss, um einen Messwert zu erhalten, statt einfach die Summe weniger Variablen zu bilden.

Die vorliegende Arbeit beschäftigt sich vorwiegend mit solchen Methoden der Testkonstruktion, bei denen auf eine Gewichtung der Items verzichtet wird. Ansätze zur Testkonstruktion mit gewichteten Variablen findet man z.B. bei Krauth (1995) oder Lord und Novick (1968). Auch Arbeiten zur Selektion von Prädiktoren in der multiplen Regression (z.B. Hocking, 1976) dürften hier relevant sein.

1.2 Definition der Gütekriterien

Wie bereits erwähnt, sind bei der Definition des Testbegriffs praktisch keine Annahmen über das Verhalten der Testperson bzw. die diesbezügliche stochastische Variable gemacht worden. Demnach kann man nahezu jede stochastische Variable als Test bezeichnen. Will man jedoch abschätzen, inwieweit aus den Testwerten Rückschlüsse auf die getesteten Personen möglich sind, so ist es notwendig anzugeben, inwieweit das Testergebnis durch Messfehler beeinflusst wurde und für welche Situationen, aufgrund des Testergebnisses, Aussagen über das Verhalten der Testperson möglich sind. In der Terminologie der Testtheorie heißt das, dass Angaben über die Reliabilität und die Validität eines Tests gemacht werden müssen.

1.2.1 Reliabilität

Die Reliabilität (Rel) eines Tests wird dabei wie folgt definiert (z.B. Steyer & Eid, 1993, S.111):

$$\mathrm{Rel}_X := \frac{\sigma^2\left(E\left[X\middle|p_u\right]\right)}{\sigma^2(X)} = \frac{\sigma^2(\tau)}{\sigma^2(X)} = 1 - \frac{\sigma^2(\varepsilon)}{\sigma^2(X)} = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\varepsilon)} = \frac{1}{1+\dfrac{\sigma^2(\varepsilon)}{\sigma^2(\tau)}} = \rho^2(X,\tau)$$

[1.2-1]

Die Reliabilität ist demnach ein spezieller Determinationskoeffizient, der den Anteil der wahren interindividuellen Unterschiede an der Testwertvarianz angibt. Je geringer die Fehlervarianz bei gegebener Varianz der wahren Werte ist, desto höher ist die Reliabilität. Die Reliabilität ist ein populationsabhängiger Kennwert der Genauigkeit eines Tests, da die Varianz der wahren Werte von der zugrundegelegten Population abhängt. Unter Umständen kann auch die Fehlervarianz populationsabhängig sein.

Betrachtet man die einzelnen Terme, welche die Reliabilität definieren, so stellt man fest, dass in jeder Definition theoretische Parameter vorkommen, die sich ohne zusätzliche Annahmen nicht

bestimmen lassen. Die Schätzung der Reliabilität erfordert zusätzliche Annahmen über die Beziehung der wahren Werte eines Tests mit den wahren Werten eines anderen Tests. So lässt sich die Reliabilität eines Tests X_1 folgendermaßen bestimmen, falls ein zweiter wesentlich τ-äquivalenter Tests X_2 bekannt ist, dessen Fehler nicht mit den Fehlern des ersten Tests korrelieren (vgl. [1.1-10] auf S. 7, [1.1-16] auf S. 10 und [1.2-1] auf S. 14 oder auch Steyer & Eid, 1993, S. 164):

$$\begin{aligned}\mathrm{Rel}(X_1) &= \frac{\sigma(\tau_1, \tau_2 - \lambda_{12})}{\sigma^2(X_1)} \\ &= \frac{\sigma(X_1, X_2)}{\sigma^2(X_1)}\end{aligned}$$

[1.2-2]

Haben die Tests außerdem die gleiche Varianz, so ist die Reliabilität die Korrelation der beiden Tests. Der üblichen Praxis, die Reliabilität eines Tests zu bestimmen, indem man die Korrelation der Testwerte bei wiederholter Testdurchführung berechnet, liegt also implizit die Vorstellung zugrunde, Test und Retest seien wesentlich τ-äquivalent mit unkorrelierten, varianzhomogenen Fehlern.

Für die Reliabilität eines Tests X_1, der bezüglich zweier anderer Tests X_2, X_3 τ-kongenerisch ist, gilt bei paarweise unkorrelierten Messfehlern (vgl. [1.1-10] auf S. 7, [1.1-17] auf S. 10 und [1.2-1] auf S. 14 oder auch Steyer & Eid, 1993, S. 211):

$$\begin{aligned}\mathrm{Rel}(X_1) &= \frac{\sigma^2(\tau_1)}{\sigma^2(X_1)} \cdot \frac{\sigma(\tau_2)\cdot\sigma(\tau_3)}{\sigma(\tau_2)\cdot\sigma(\tau_3)} \\ &= \frac{\sigma(\tau_1, \tau_2)\cdot\sigma(\tau_1, \tau_3)}{\sigma^2(X_1)\cdot\sigma(\tau_2, \tau_3)} \\ &= \frac{\sigma(X_1, X_2)\cdot\sigma(X_1, X_3)}{\sigma^2(X_1)\cdot\sigma(X_2, X_3)} \\ &= \frac{\rho(X_1, X_2)\,\rho(X_1, X_3)}{\rho(X_2, X_3)}\end{aligned}$$

[1.2-3]

Bei τ-kongenerischen Tests sind also drei Messungen notwendig, um die Reliabilität zu bestimmen.

Wenn ein Test T sich additiv in n lokal unkorrelierte Tests X_1, ... , X_n zerlegen lässt, so sind sowohl Cronbachs-α als auch der λ_2-Koeffizient von Guttmann (1945) grundsätzlich untere

Schranken der Reliabilität. Der λ_2-Koeffizient ist immer eine mindestens genauso gute untere Schranke wie Cronbachs-α (Krauth, 1995, S. 259):

$$\mathrm{Rel}(T) \geq \lambda_2 := 1 - \frac{\sum_{i=1}^{n} \sigma^2(X_i) + \frac{2n}{n-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\sigma(X_i, X_j)\right)^2}{\sigma^2(T)} \geq \alpha := \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} \sigma^2(X_i)}{\sigma^2(T)}\right)$$

[1.2-4]

Nur für den Fall, dass der Test *T* die Summe wesentlich τ-äquivalenter Tests ist, gilt (Krauth, 1995, S. 260):

$$\mathrm{Rel}(T) = \lambda_2 = \alpha$$

[1.2-5]

Aus den obigen Ausführungen sollte man jedoch keinesfalls den Schluss ziehen, dass die Reliabilität eines Tests immer größer sein muss als empirische Schätzungen von α und λ_2. Schätzungen dieser Koeffizienten können aufgrund von zufälligen Abweichungen durchaus größer sein als die Reliabilität. (Es kann auch vorkommen, dass die Koeffizienten oder ihre Schätzer negativ sind; Krauth, 1995). Dennoch empfiehlt Krauth (1995, S. 256), alle möglichen additiven Zerlegungen des Tests vorzunehmen und die höchsten resultierenden Wert von α bzw. λ_2 als Schätzer der Reliabilität zu verwenden. Schon bei relativ kurzen Tests sind bei diesem Verfahren unzähligen Schätzungen vorzunehmen. Da jede dieser Schätzungen mit einem Messfehler behaftet ist, besteht daher die Gefahr, dass man additive Zerlegungen des Tests wählt, bei denen die in der Stichprobe ermittelten Koeffizienten die entsprechenden Populationskennwerte stark überschätzen (→ Regression zur Mitte). Dies kann (auch bei τ-äquivalenten Teiltests) zu inflationären Schätzungen der Reliabilität führen. In jedem Fall kann man bei der Methode von Krauth (1995) nicht mehr davon ausgehen, konservative, d.h. tendenziell zu niedrige, Schätzungen der Reliabilität zu erhalten.

Auch andere Methoden zur Bestimmung der Reliabilität (split-half, varianzanalytische Bestimmung, Kuder-Richardson Formeln) gehen in der Regel davon aus, dass die Items eines Tests essentiell τ-äquivalent sind. Ist diese Annahme nicht erfüllt, so führen die verschiedenen Verfahren zur Bestimmung der Reliabilität nicht nur aufgrund von Schätzfehlern u.U. nicht mehr zu denselben Werten, sondern auch deswegen, weil die entsprechenden Kennwerte (Cronbachs-α, λ_2, split-half Reliabilität etc.) in der Population unterschiedlich sind. Daraus sollte man jedoch keinesfalls die Folgerung ableiten, dass es verschiedene Arten der Reliabilität gäbe,

wie einige Texte zur Testtheorie nahe legen. Die mangelnde Übereinstimmung der verschiedenen Schätzer der Reliabilität ist vielmehr darauf zurückzuführen, dass die Voraussetzungen für ihre Anwendungen nicht immer gegeben sind. So sollte beispielsweise eine Retest-Korrelation nicht als Reliabilitätskoeffizient interpretiert werden, wenn Test und Retest nicht essentiell τ-äquivalent sind. Bei lediglich τ-kongenerischen Messungen braucht man beispielsweise mindestens drei Messzeitpunkte um die Reliabilität anhand von [1.2-3] (S. 15) zu bestimmen. [1.2-1] (S. 14) und [1.2-2] (S. 15)sind in diesem Fall keine angemessenen Schätzer für die Reliabilität.

1.2.2 Validität

Die Validität oder Gültigkeit eines Tests gibt an, inwieweit "ein Test, dasjenige Persönlichkeits- oder Verhaltensmerkmal misst, das er messen soll oder zu erfassen vorgibt" (Amelang & Zielinski, 1997, S. 155). Die Validität eines Tests lässt sich nicht so leicht beurteilen wie dessen Reliabilität. Dies hängt nicht zuletzt damit zusammen, dass man je nachdem, unter welchem Gesichtspunkt man die Validität eines Tests beurteilt, zu völlig unterschiedlichen Feststellungen gelangen kann. Es gibt mehrere Möglichkeiten, die verschiedenen Ansätze zur Beurteilung der Validität zu klassifizieren. Grob lassen drei Aspekte der Validität voneinander abgrenzen.

Interne Validität

Die internale oder interne Validität eines Tests gilt dann als gegeben, wenn sich die Annahmen über das Antwortverhalten der Personen in der Testsituation anhand empirischer Daten bestätigen lassen (Rost, 1996). Die Annahmen über das Antwortverhalten beziehen sich hierbei weniger auf die zu messende Eigenschaft, also den Gegenstand der Messung, als vielmehr auf das Testmodell, das bei der Testkonstruktion zugrunde gelegt wird. Überprüft wird die interne Validität z.B. anhand von Modellgeltungstests, die entweder quantitative Maße für die Güte der Modellanpassung liefern oder einen statistischen Test mit dem Testmodell als Nullhypothese darstellen. Manchmal wird die Modellanpassung eines Testmodells anhand deskriptiver Werte eingeschätzt, die keine direkte quantitative Aussage erlauben, etwa der Scree-Test der Faktorenanalyse zur Überprüfung der Eindimensionalität eines Itempools.

Wenngleich sich Aussagen zur internen Validität zunächst nur auf die Angemessenheit eines formalen Modells zur Beschreibung des Verhaltens in der Testsituation beziehen und zumindest im formallogischen Sinn keinen induktiven Schluss auf den Gegenstand der Messung erlauben, so können sie doch der Ausgangspunkt für die Modellbildung sein. Zeigt sich etwa, dass sich die Annahme der Eindimensionalität eines Itempools zur Erfassung der Aggressivität nicht halten

lässt, so wird man sich überlegen, ob es nicht angemessener ist, Aggressivität als mehrdimensionales Konstrukt aufzufassen. Gigerenzer (1981) macht darauf aufmerksam, dass bereits die Wahl einer Methode zur Analyse eines Gegenstandsbereichs eine modellbildende Funktion für den Gegenstand hat.

Externe Validität

Der Begriff der externalen oder externen Validität bezieht sich auf den statistischen Zusammenhang des Tests mit einem Außenkriterium. Als Kriterium sind vor allem solche Variablen geeignet, die auch als Messwerte für die betreffende Eigenschaft in Frage kommen, und die einen höheren Status als der zu validierende Test haben. Burisch (1984a) nennt solche Variablen echte Kriterien und grenzt sie von Quasi-Kriterien ab, die ein geringeres Ansehen als der zu validierende Test haben. Ist der einzige Zweck eines Tests, Informationen über die Werte einer anderen Variable zu gewinnen (beispielsweise den Ausbruch einer Krankheit oder das Suizidrisiko), so nennt Burisch (1984a) die externe Variable Target-Variable. Wird die Kriteriumsvariable zu einem späteren Zeitpunkt als der zu validierende Test gemessen, so spricht man von prognostischer oder prädiktiver Validität, bei nahezu zeitgleicher Messung von Test und Kriterium von konkurrenter Validität.

Da häufig verschiedene Variablen als Kriterien in Frage kommen, kann ein Test durchaus eine Vielzahl differierender Kennwerte für seine externe Validität haben. Die Höhe der Korrelation zwischen Test X und Kriterium C hängt dabei nicht nur vom Grad der konzeptuellen Gemeinsamkeit ab, sondern auch von der Reliabilität von Test und Kriterium (Amelang & Zielinski, 1997, S. 159) Die Minderung der Korrelation durch Messfehler lässt sich mit Hilfe von sogenannten Verdünnungsformeln beschreiben (z.B. Amelang & Zielinski, 1997 oder auch [1.2-7] auf S. 19 mit $\rho(\varepsilon_C, \varepsilon_X) = 0$):

$$\rho_{C,X} = \rho(C, \tau_X) \cdot \sqrt{\mathrm{Rel}_X} = \rho(\tau_C, X) \cdot \sqrt{\mathrm{Rel}_C} = \rho(\tau_C, \tau_X) \cdot \sqrt{\mathrm{Rel}_C \cdot \mathrm{Rel}_X} \qquad \textbf{[1.2-6]}$$

Aus diesen Formeln lassen sich Minderungskorrekturen ermitteln, die angeben, wie stark der statistische Zusammenhang der beiden Variablen wäre, wenn ihre Messung fehlerfrei wäre.

[1.2-6] gilt jedoch nur, wenn die Fehler von Test und Kriterium nicht korrelieren. Während die Annahme unkorrelierter Messfehler zwischen den Items eines Tests kaum zu rechtfertigen ist (wegen reaktionskontingenten Übungseffekten, Einflüssen von States bei Traittests, Ermüdung, Tendenzen zur konsistenten Selbstdarstellung), ist es aufgrund der unterschiedlichen Herkunft

der Daten häufig durchaus angemessen zu postulieren, dass die Fehler (der Items) eines Tests nicht mit den Fehlern des Kriteriums korrelieren. Die Verdünnungsformeln lassen sich jedoch so formulieren, dass sie auch bei korrelierten Fehlern gelten (vgl. [1.1-10] auf S. 7 und [1.2-1] auf S. 14 oder auch Zimmermann & Williams, 1980):

$$\begin{aligned}
\rho_{C,X} &= \frac{\sigma(\tau_C,\tau_X)}{\sigma_C\cdot\sigma_X} + \frac{\sigma(\varepsilon_C,\varepsilon_X)}{\sigma_C\cdot\sigma_X} \\
&= \frac{\rho(\tau_C,\tau_X)\sqrt{\sigma_C^2\cdot\sigma_X^2\cdot \mathrm{Rel}_C\cdot \mathrm{Rel}_X}}{\sigma_C\cdot\sigma_X} + \frac{\rho(\varepsilon_C,\varepsilon_X)\sqrt{\sigma_C^2\cdot\sigma_X^2\cdot(1-\mathrm{Rel}_C)\cdot(1-\mathrm{Rel}_X)}}{\sigma_C\cdot\sigma_X} \\
&= \rho(\tau_C,\tau_X)\cdot\sqrt{\mathrm{Rel}_C\cdot\mathrm{Rel}_X} + \rho(\varepsilon_C,\varepsilon_X)\cdot\sqrt{(1-\mathrm{Rel}_C)\cdot(1-\mathrm{Rel}_X)} \\
&= \rho(C,\tau_X)\cdot\sqrt{\mathrm{Rel}_X} + \rho(\varepsilon_C,\varepsilon_X)\cdot\sqrt{(1-\mathrm{Rel}_C)\cdot(1-\mathrm{Rel}_X)} \\
&= \rho(\tau_C,X)\cdot\sqrt{\mathrm{Rel}_C} + \rho(\varepsilon_C,\varepsilon_X)\cdot\sqrt{(1-\mathrm{Rel}_C)\cdot(1-\mathrm{Rel}_X)}
\end{aligned}$$

[1.2-7]

Inhaltliche Validität

Die inhaltliche Validität eines Tests ist dann gegeben, wenn man aufgrund von theoretischen Erwägungen, die nicht durch empirische Daten gestützt sind, zu der Überzeugung gelangt, dass ein Test ein bestimmtes Merkmal misst. Mit theoretischen Erwägungen sind in diesem Zusammenhang sowohl wissenschaftliche Theorien als auch subjektive oder implizite Theorien über den Gegenstand der Messung gemeint. Diese mehr oder weniger explizierten theoretischen Vorstellungen beziehen sich dabei sowohl auf die zu messende Eigenschaft als auch auf den Prozess, durch den sich die Eigenschaft bei der Messung manifestiert. Wird die inhaltliche Validität eines Tests ohne eingehende theoretische Begründung beurteilt, so spricht man auch von Augenscheinvalidität oder intuitiver Validität. Prüft man beispielsweise die Eignung einer Sekretärin, indem man die Anzahl der Tipp- und Schreibfehler in einem fremdsprachigen Diktat bestimmt, so bedarf es keiner ausgereiften Theorie, um zu beurteilen, ob dieser Test etwas über die Fähigkeit zur Ausübung dieses Berufes aussagt. Ist die Entsprechung von Test und zu messender Eigenschaft weniger offensichtlich, so kann man die inhaltliche Validität eines Tests bzw. einzelner Items durch mehrere Personen einschätzen lassen und quantitative Maße für die inhaltliche Validität bestimmen (Amelang & Zielinski, 1997).

Beurteilt man die inhaltliche Validität eines Tests dagegen aufgrund eingehender theoretischer – sachlogischer und begrifflicher – Erwägungen, so spricht man von Konstruktvalidität. Bei der Beurteilung der Konstruktvalidität schließt man jedoch im Allgemeinen auch die Ergebnisse empirischer Untersuchungen mit ein. Dabei kommen sowohl quasi-experimentelle

Untersuchungen in Frage, bei denen die Testwerte als unabhängige Variable dienen, als auch korrelative Studien, welche die Zusammenhänge mit anderen externen Variablen untersuchen. Aus den theoretischen Vorstellungen über die durch einen Test gemessene Variable werden Hypothesen über die Ergebnisse dieser Untersuchungen abgeleitet. Je besser die empirischen Ergebnisse mit den Hypothesen übereinstimmen, desto günstiger wird man die Konstruktvalidität eines Tests beurteilen. Zeigt ein Test entsprechend den theoretischen Erwartungen hohe Korrelationen mit anderen Tests oder Variablen, so hat er eine hohe konvergente Validität. Sind weiterhin die Korrelationen mit anderen Tests erwartungsgemäß niedrig, so hat er eine hohe diskriminante Validität. Beim Multi-Trait-Multi-Method Ansatz (Campell & Fiske, 1959) wird systematisch die konvergente und diskriminante Validität eines Tests untersucht. Dabei wird auch dem Umstand Rechnung getragen, dass Gemeinsamkeiten in der Erhebungsmethode – unabhängig von etwaigen konzeptuellen Beziehungen – Korrelationen zweier Tests bedingen können.

Des Weiteren dürfte auch die interne Validität bei der Beurteilung der Konstruktvalidität eines Tests eine gewisse Rolle spielen, da man nur dann sicher sein kann, dass der Test ein Konstrukt in sinnvoller Weise erfasst, wenn es auch gelingt, ein adäquates formales Modell zu entwickeln, das das Verhalten in der Testsituation beschreibt. Das Konzept der Konstruktvalidität umfasst somit alle anderen Aspekte der Validität. Die Konstruktvalidität eines Tests lässt sich daher auch meist nicht abschließend beurteilen. Vielmehr wird versucht, unter Berücksichtigung sowohl aller theoretischer Aspekte als auch der empirischen Datenlage zu beurteilen, inwieweit ein Test dem Anspruch gerecht wird, ein genau umschriebenes Konstrukt in sinnvoller Weise zu erfassen. Die Sicherung der Konstruktvalidität stellt eher einen fortlaufenden Prozess dar und lässt sich im Allgemeinen nicht durch einzelne numerische Werte kennzeichnen.

Bereits aus [1.2-6] (S. 18) lässt sich ableiten, dass die Validität der Reliabilität als Gütekriterium übergeordnet ist. Eine hohe Validität ist nämlich ein hinreichendes Kriterium für die Reliabilität eines Tests, da die Reliabilität des Tests mindestens so groß sein muss wie die quadrierte Validität. Eine hohe Reliabilität ist dagegen kein Garant für die Validität eines Tests. Diese Aussage gilt jedoch nur dann uneingeschränkt, wenn die Messfehler von Test und Kriterium nicht korrelieren (vgl. [1.2-7] auf S. 19) und das zu messende Persönlichkeitsmerkmal a-priori feststeht. Bei einer operationalen Definition von Persönlichkeitsmerkmalen ist die Reliabilität eines Tests per definitionem gleich der Validität. Operationale Definitionen sind jedoch wissenschaftstheoretisch sehr problematisch (Westermann, 1990). Auch von einem pragmatischen Standpunkt aus betrachtet sind operationale Definitionen unbefriedigend, wenn es

nicht gelingt die inhaltliche Bedeutung des erhobenen Merkmals zu klären. Selbst die Vorhersage von Targetvariablen (s.o.) im Rahmen von multiplen Regressionen dürfte sonst in der Regel von einem fragwürdigen Nutzen sein, da man die Natur der gefundenen Zusammenhänge nicht versteht.

Wenn in den folgenden Kapiteln von Validität gesprochen wird, so ist in der Regel die externe Validität gemeint, auch wenn dies nicht mehr explizit betont wird. Allerdings muss das Außenkriterium dabei nicht unbedingt eine konkrete empirische Variable sein, sondern die dargestellten Überlegungen gelten auch für hypothetische Größen, wie etwa die „wahren" Ausprägungen der Personen auf den intendierten Persönlichkeitsdimensionen. Selbst wenn es nicht gelingt, eine konkrete empirische Variable zu finden, die dieser hypothetische Variablen entspricht, lassen sich dennoch Aussagen über die Validität hinsichtlich dieser Variablen machen (z.B. [1.2-6] auf S. 18).

1.3 Testlänge und Gütekriterien

Dass eine Verlängerung von Tests durch Hinzufügen weiterer Items zu einer Verbesserung der Gütekriterien führt, halten die meisten Sozialwissenschaftler für eine testtheoretische Binsenweisheit[2]. Laut Lienert und Raatz (1994) gilt die Verlängerung von Tests gar als „via regia der Reliabilitätsverbesserung" (S. 209). Diese Auffassung stützt sich v.a. auf die Spearman-Brown Formel für die Reliabilität (vgl. Steyer & Eid, 1993, S. 165)

$$\mathrm{Rel_T} = \frac{n \cdot \mathrm{Rel_X}}{1 + (n-1) \cdot \mathrm{Rel_X}} = \frac{\mathrm{Rel_X}}{\mathrm{Rel_X} + (1 - \mathrm{Rel_X}) \cdot n^{-1}} \xrightarrow{n \to \infty} 1$$

[1.3-1]

und die Validität (vgl. Lienert & Raatz, 1994, S. 255)

$$\rho(\mathrm{T,C}) = \frac{\rho(\mathrm{X,C})}{\sqrt{\mathrm{Rel_X} + \frac{(1 - \mathrm{Rel_X})}{n}}} \xrightarrow{n \to \infty} \frac{\rho(\mathrm{X,C})}{\sqrt{\mathrm{Rel_X}}}$$

[1.3-2]

[1.3-1] und [1.3-2] beschreiben den Zusammenhang der Testlänge mit der Reliabilität und Validität eines Tests T in Abhängigkeit der Reliabilität der Items ($\mathrm{Rel_X}$). Reliabilität und

[2] Es gibt sogar Ansätze, Maßnahmen zur Verbesserung der Gütekriterien eines Tests anhand eines Kennwerts (coefficient of effective length) zu beurteilen, der angibt, welche Veränderungen in der Testlänge notwendig gewesen wären, um die entsprechenden Veränderungen in der Reliabilität und Validität zu erreichen (z.B. Edwards, 1981).

Validität steigen demnach mit zunehmender Testlänge streng monoton an. Die Reliabilität konvergiert mit zunehmender Anzahl der Items gegen eins, während der Grenzwert der Validität die minderungskorrigierte Itemvalidität (vgl. [1.2-6] auf S. 18) ist. [1.3-1] und [1.3-2] gelten jedoch nur, wenn folgende Voraussetzungen erfüllt sind:

- Essentielle τ-Äquivalenz ([1.1-16] auf S. 10)
- Varianzhomogenität der Items in Population
- Unkorreliertheit der Messfehler der Items untereinander sowie mit dem Messfehler des Kriteriums

Verletzungen der τ-Äquivalenz oder der Varianzhomogenität der Items können dazu führen, dass Reliabilität und Validität nicht mehr wie in [1.3-1] und [1.3-2] monoton mit der Testlänge ansteigen (Yousfi, 2004a). Bei der Auswahl von Items aus einem infiniten Itempool erreichen Reliabilität und Validität dennoch schließlich dieselben Grenzwerte, wie bei Geltung von [1.3-1] und [1.3-2]. Bei (positiven) Korrelationen zwischen den Messfehlern der Items ist der Grenzwert der Reliabilität dagegen grundsätzlich kleiner als bei Geltung von [1.3-1] und [1.3-2]. In diesem Fall ist es sogar möglich, dass die Reliabilität mit zunehmender Testlänge streng monoton fällt (Yousfi, 2004a).

Bei τ-kongenerischen Items (vgl. [1.1-17] auf S. 10), deren Messfehler mit dem Messfehler des Kriteriums unkorreliert sind, ist es nicht möglich, dass sich die Reliabilität und Validität mit zunehmender Testlänge unterschiedlich entwickeln (vgl. [1.2-7] auf S. 19). Der Zusammenhang der Validität mit der Testlänge lässt sich dann aus dem Zusammenhang zwischen der Reliabilität und Testlänge herleiten (Yousfi, 2004a). Bei Verletzungen der τ-Kongenerität (Loevinger, 1954) und bei Korrelationen zwischen den Messfehlern der Items und des Kriteriums (Williams & Zimmerman, 1982) muss eine Veränderung der Reliabilität dagegen nicht mit einer gleichgerichteten Veränderung der Validität einhergehen (Verdünnungsparadox).

Dass ein negativer Zusammenhang zwischen der Testlänge und den Gütekriterien möglich ist, lässt sich nicht nur analytisch zeigen (Loevinger, 1954; Zimmerman & Williams, 1980; Ulrich, 1985; Yousfi, 2004a), sondern es gibt auch zahlreiche empirische Beispiele (z.B. Bell & Lumsden, 1980, Broughton, 1984, Burisch, 1984a, 1984b, 1997). Paunonen und Jackson (1985) sind allerdings der Auffassung, dass sich nur bei gezielter Auswahl der Items ein negativer Zusammenhang zwischen der Testlänge und der Testgüte ergeben kann. Yousfi (2004a) zeigt

jedoch, dass dies selbst bei zufälliger Auswahl aus einem *finiten* τ-äquivalenten Itempool mit unkorrelierten Messfehlern möglich ist.

2 Itemselektion in der klassischen Testtheorie

2.1 Selektion anhand von Itemkennwerten

Die Itemselektion erfolgt im Rahmen der klassischen Testtheorie anhand verschiedener Kennwerte. Im Idealfall sind die Validitäten der einzelnen Items bekannt. Häufig haben die verfügbaren Außenkriterien jedoch nicht unbedingt den Status eines echten Außenkriteriums im Sinne von Burisch (1984a), d.h. im Zweifel würde man ihnen nicht unbedingt mehr Vertrauen schenken als den Testitems selbst. Stattdessen stellen diese Kriterien eher Zielvariablen dar, hinsichtlich derer die Items und der resultierende Test lediglich konvergente Validität haben sollten. Der externalen Strategie der Skalenkonstruktion wird zudem nachgesagt, dass sie zu inhomogenen Tests führt (Amelang & Zielinski, 1997), was die Einschätzung der psychologischen Bedeutung einer Skala weiter erschweren kann. Nicht selten werden Tests daher ganz ohne jegliche Information über die empirische Validität der Items konstruiert. Bei der rationalen Strategie der Skalenkonstruktion, die jedoch nicht als Methode der klassischen Testtheorie betrachtet werden kann, verlässt man sich allein auf Intuition und inhaltlich-theoretische Erwägungen. Hier stehen also die Inhalts- oder Augenscheinvalidität der Items im Zentrum der Analyse. Bei der induktiven Strategie bilden dagegen die statistischen Zusammenhänge zwischen den Items die Grundlage der Skalenkonstruktion. Neben der Faktorenanalyse sind Itemanalysen nach klassischer oder probabilistischer Methodik den Methoden der induktiven Skalenkonstruktion zuzurechnen.

Bei faktorenanalytischer sowie klassischer Methodik der Testkonstruktion werden meist Korrelationen der Items mit dem Skalenwert als Selektionskriterium verwendet. Wenn die Skalenwerte die Summe der einzelnen Items sind, nennt man die Korrelationen mit der Skala Trennschärfe. Wenn die Skala dagegen auf faktorenanalytischem Wege gewonnen worden werden, so nennt man die Korrelation des Faktors mit den Items Faktorladungen. Für die endgültige Skala werden meist einfach diejenigen Items aufsummiert, die hinreichend hohe Trennschärfen oder Faktorladungen erreichen. Dieses Vorgehen soll die Eindimensionalität der resultierenden Skala sichern, d.h. es soll sichergestellt werden, dass alle Items dasselbe Merkmal erfassen. Die Korrelation mit dem Gesamtskalenwert hängt jedoch nicht nur von der Affinität zu dem erfassten Merkmal ab, sondern auch vom Schwierigkeitsindex des Items oder präziser von der Randverteilung des Items. Je mehr sich die Randverteilungen von Item und Test unterscheiden, desto geringer ist die Trennschärfe, die maximal erreicht werden kann (Lord & Novick, 1968). Es sind verschiedene Methoden vorgeschlagen worden, die diesem Umstand Rechnung tragen sollen:

- Ein informelles, häufig empfohlenes Vorgehen besteht darin, Items mit niedriger Trennschärfe nur dann auszuscheiden, wenn sie einen mittleren Schwierigkeitsindex haben.

- Eine standardisierte Variante dieses Vorgehens besteht darin, Items anhand des sogenannten Selektionskennwerts auszuwählen. Der Selektionskennwert ist der Quotient aus der Trennschärfe und der doppelten Standardabweichung des Items. Da die Standardabweichung umso geringer ist, je extremer der Schwierigkeitsindex ist, werden hier auch Items mit geringerer Trennschärfe gewählt, wenn sie einen extremen Schwierigkeitsindex haben (Lienert & Raatz, 1994).

- Moosbrugger und Zistler (1993) schlagen vor, die Trennschärfe zu berechnen, indem man für jedes Item zunächst die Stichprobe der Personen nach Maßgabe ihrer Skalenwerte proportional zum Schwierigkeitskoeffizienten aufteilt, und anschließend mithilfe des Phi-Koeffizienten ermittelt, inwieweit die Gruppe mit höheren Skalenwerten das Item häufiger gelöst hat. Da die Personenstichprobe so eingeteilt wurden, dass die Randverteilungen von Itembeantwortung und Gruppenzugehörigkeit identisch sind, kann der so ermittelte Phi-Koeffizient auch Werte von 1 annehmen. Im Gegensatz zu dem Verfahren von Moosbrugger und Zistler (1993) bietet die Berechnung des Phi-Koeffizienten über medianhalbierte Skalenwerte bei variierenden Schwierigkeitsparametern grundsätzlich keine Vorteile gegenüber der Berechnung von Produkt-Moment oder punkt-biserialen Korrelationen.

- Eine weitere Alternative ist die Berechnung von biserialen, rangbiserialen, polyserialen und polychorischen Korrelationen. Diese Maße versuchen den „wahren“ Zusammenhang zwischen den Skalenwerten und der Itembeantwortung abzuschätzen. Sie gehen davon aus, dass dem Item (und bei polychorischen oder rangbiserialen Korrelation auch dem Test) eine normalvereilte kontinuierliche Variable zugrunde liegt, die nur durch das Antwortformat der Items künstlich in geordnete Kategorien aufgespaltet wird.

Die Abhängigkeit der Trennschärfe von der Itemschwierigkeit wird in jedem Lehrbuch zur klassischen Testtheorie ausführlich behandelt. Erstaunlicherweise bleibt die Frage, ob die Trennschärfe prinzipiell überhaupt dazu geeignet ist, die Testgüte zu sichern, meist völlig ausgeklammert. Aus der testtheoretischen Literatur lässt sich lediglich entnehmen, dass hohe Itemvaliditäten hohe Itemtrennschärfen implizieren, während hohe Itemtrennschärfen ihrerseits nur bei einer hohen Testreliabilität möglich sind (Lord & Novick, 1968, S.331-333). Wie sich

die Itemselektion anhand der Trennschärfe dagegen auf die Reliabilität und Validität eines Tests auswirkt, ist eine weithin ungeklärte Frage (Krauth, 1995, S.278).

Yousfi (2004b) hat den Zusammenhangs der Trennschärfe mit den Gütekriterien eingehend analysiert und ist dabei zu folgenden Ergebnissen gekommen:

- Parallele Items haben identische Trennschärfen. Bei parallelen Items sind Trennschärfen als Hilfsmittel für die Itemselektion daher überflüssig.

- Bei τ-äquivalenten Items geht eine höhere Trennschärfe grundsätzlich mit einer höheren Itemvalidität einher. Die Präferenz für trennscharfe Items führt bei τ-Äquivalenz zu Tests mit maximaler Reliabilität und Validität.

- Bei τ-kongenerischen Items geht eine höhere korrigierte Trennschärfe nicht unbedingt mit einer höheren Itemvalidität einher. Auch für die Reliabilität und Validität des Tests kann es günstiger sein, ein Item mit einer geringeren Trennschärfe zu bevorzugen.

- Wenn man keine Annahmen über den Zusammenhang zwischen den wahren Werten der Items macht, ist die Trennschärfe in der Regel ein guter Indikator der Itemreliabilität. Die Präferenz für trennscharfe Items führt zu Tests mit hoher Reliabilität und interner Konsistenz.

- Wenn man keine Annahmen über den Zusammenhang zwischen den wahren Werten der Items macht, so ist die Testvalidität ein Moderator des Zusammenhangs zwischen Trennschärfe und Itemvalidität. Je geringer die Reliabilität des Tests ist desto stärker ist Moderatorwirkung der Testvalidität. Nur bei validen Tests ist die Trennschärfe demnach ein guter Indikator der Itemvalidität.

- Bei gleicher Itemvalidität führt die Präferenz für trennscharfe Items zu Tests geringerer Validität (vgl. auch Ulrich, 1985).

- Da die Trennschärfe einerseits bei validen Tests ein Indikator der Itemvalidität ist und anderseits die Präferenz für trennscharfe Items bei gleicher Itemvalidität ungünstig für die Validität des Tests ist, dürfte die Elimination wenig trennscharfer Items allenfalls bei wenig reliablen, aber dennoch sehr validen Tests zu wesentlich besseren Ergebnissen führen als die Zufallsauswahl. Da die Validität jedoch nicht größer sein kann als Wurzel

aus der Reliabilität (vgl. [1.2-6] auf S. 18) dürften derartige Konstellation in der Praxis nur selten vorkommen.

2.2 Itemselektion anhand von Skalenkennwerten

2.2.1 Theoretische Begründung

Aus dem Arbeitskreis von Cattell wurde bereits in den sechziger Jahren massive Kritik an der gängigen Skalenkonstruktion mit den Methoden der klassischen Testtheorie geäußert (Cattell & Radcliffe, 1962; Cattell und Tsujioka; 1964). Kritisiert wurde, dass diese herkömmlichen Methoden sich fast ausschließlich auf die Sicherung der Homogenität einer Skala konzentrieren. Da natürliches Verhalten in der Regel von einer Vielzahl von Persönlichkeitsfaktoren abhänge, sei es jedoch außerordentlich schwierig, eine hinreichende Zahl von Items zu finden, die vorwiegend den erwünschten Persönlichkeitsfaktor erfassen. Geometrisch lässt sich diese Aussage veranschaulichen, indem man sich die möglichen Items als Vektoren in einem mehrdimensionalen Raum von Persönlichkeitsfaktoren vorstellt. Da es in diesem Raum sehr viele Richtungen gibt, in welche die Vektoren weisen können, sei es nahezu unmöglich, Vektoren zu finden, die sich in ihrer Ausrichtung kaum unterscheiden. Um dennoch homogene Skalen zu erreichen, würden meist Items generiert, die sich inhaltlich so ähnlich seien, dass sie nicht nur dieselben Persönlichkeitsfaktoren abbilden, sondern vorwiegend dieselben spezifischen Varianzanteile („bloated specific“ disorder), die eigentlich nicht Gegenstand der Messung sind. Obwohl die Items jeweils ein ähnliches Gemenge von Persönlichkeitsfaktoren erfassen, gelingt es meist nicht, ausschließlich den Persönlichkeitsfaktor abzubilden, der gemessen werden soll. Die Items haben also ein homogenes Ladungsmuster auf einer Vielzahl von Persönlichkeitsfaktoren. Die resultierenden Skalen sind dann zwar homogen, aber sie sind systematisch mit spezifischer Varianz sowie konfundierenden Persönlichkeitsfaktoren verunreinigt. Daher haben sie zwar u.U. eine hohe Reliabilität, aber auf Kosten einer sehr geringen Validität.

Dieser Tradition der Skalenkonstruktion („itemmetrics“) stellt Cattell seine strukturelle Methode der Skalenkonstruktion gegenüber, die auf dem faktorenanalytisch gewonnenen Wissen über die Struktur der Persönlichkeit aufbaut. Ziel ist nicht mehr die Erstellung von homogenen Skalen, sondern von gepufferten Skalen, bei denen die Einflüsse von konfundierenden Faktoren über Suppressionseffekte eliminiert werden. Cattell und Tsujioka (1964) entwickeln eine Reihe von Methoden, mit denen die klassischen sowie neuere Gütekriterien im Rahmen dieses Ansatzes bewerten werden können. Sie geben jedoch keinen Algorithmus an, der die Pufferung von

Skalen sichert. Sie empfehlen lediglich, darauf zu achten, dass die Ladungen (genauer: die standardisierten Regressionskoeffizienten) der Items auf den fremden Faktoren ausbalanciert sind. Dieses Vorgehen zielt jedoch nicht direkt auf eine Sicherung der faktoriellen Validität der Skala ab, sondern auf eine hohe „factor-trueness“. Damit ist die minderungskorrigierte Korrelation mit dem intendierten Faktor gemeint, bei welcher der Einfluss der spezifischen Varianzanteile herausgerechnet wird. Bartlett (1937) sowie Anderson und Rubin (1956) geben Algorithmen an, bei denen die Faktortreue einer Skala durch eine entsprechende Gewichtung der Items erreicht wird.

Bei dem von Cattell entwickelten 16 PF werden jedoch keine Gewichtungen der Items vorgenommen. Eine Vielzahl von Untersuchungen (z.B. Howarth, Browne & Marceau, 1972; Saville & Blinkhorn, 1981; Greif, 1970; Meyer, Arnold, Freitag & Balck, 1977) zum 16 PF belegen jedoch, dass weder klassische Gütekriterien noch die von Cattell aufgestellten neuen Gütekriterien vom 16 PF in hinreichendem Maße erfüllt werden. Dies dürfte denn auch der Hauptgrund dafür sein, dass die von Cattell vorgeschlagenen Prinzipien der Skalenkonstruktion in der Praxis kaum beachtet werden[3]. Die Überprüfung der Gütekriterien des 16 PF ersetzt jedoch nicht die inhaltliche Auseinandersetzung mit der Argumentation von Cattell. Daher sollen im Folgenden, neben Cattells Ansatz, auch andere damit verwandte multivariate Skalenkonstruktionsmethoden besprochen und empirisch untersucht werden.

2.2.2 Beschreibung der Methoden

Wenn man eine Gewichtung der einzelnen Items zulässt, dann stellt die Entwicklung einer faktorreinen Skala bei bekannten Faktorladungen kein Problem dar[4]. In diesem Fall lassen sich nämlich Faktorwerte schätzen, die nicht von den Werten auf anderen gemeinsamen Faktoren beeinflusst werden[5] (Bartlett; 1937; Anderson & Rubin, 1956). Verzichtet man jedoch auf eine Gewichtung der Items, so ist es im Allgemeinen nicht möglich, absolut faktorreine Skalen zu erstellen (vgl. die Erläuterungen zu [1.1-21] auf S. 13).

---

[3] Lediglich der Berliner Intelligenztest von Jäger, Süss und Beauducel (1997) orientiert sich explizit an Ideen von Cattell.

[4] Voraussetzung ist jedoch, dass die Dimension des von den Ladungsvektoren aufgespannte Vektorraums nicht niedriger ist als die Anzahl der Persönlichkeitsfaktoren.

[5] Da die wahren Faktorladungen jedoch nicht bekannt sind und an einer Stichprobe geschätzt werden, sind jedoch auch die Faktorwerte nicht völlig unabhängig von den Ausprägungen auf anderen Faktoren.

Dennoch könnte man versuchen, die Items so zusammenzustellen, dass die Verunreinigung durch fremde Faktoren möglichst klein ist. Wenn Verunreinigungen durch Messfehler (und andere spezifische Varianzanteile) ebenso wenig erwünscht sind wie systematische Abweichungen vom intendierten Faktor im Raum der gemeinsamen Faktoren, dann kann man einfach die Validität, d.h. die Korrelation mit dem erwünschten Faktor, maximieren. Will man dagegen nur sicherstellen, dass die Skala bis auf Messfehler (und andere spezifische Varianzanteile) mit dem intendierten Faktor übereinstimmt, so muss man die Faktortreue („factor trueness“) sensu Cattell maximieren. Die Faktortreue ist dann maximal, wenn Korrelationen der Skala mit fremden Faktoren nur etwaige Zusammenhänge zwischen den (oblique rotierten) Faktoren widerspiegeln. Faktortreue definiert Cattell als die Korrelation zwischen dem erwünschten Faktor und der Projektion der Skala in den Raum der gemeinsamen Faktoren. Die Faktortreue ist demnach die minderungskorrigierte Validität hinsichtlich des erwünschten Faktors, bei der Einflüsse spezifischer Varianzanteile herausgerechnet sind. Man definiert sie daher sinnvollerweise als den Quotienten aus der Validität und der Wurzel der Kommunalität der Skala. Cattell und Tsujioka (1964, S.20) sagen jedoch, die Validität sei das Produkt aus Faktortreue und Kommunalität. Dieser Irrtum ist darauf zurückzuführen, dass sie die Kommunalität mit der Länge Vektors gleichsetzen, der sich durch die Projektion der z-standardisierten Skala in den Raum der gemeinsamen Faktoren ergibt (Cattell & Tsujioka, 1964, S.19). Tatsächlich ist die Länge dieses Vektors jedoch die Wurzel aus der Kommunalität.

Dessen ungeachtet gilt die Behauptung von Cattell und Tsujioka, dass eine Erhöhung der Faktortreue nicht notwendigerweise zu einer verbesserten Validität führt, da hierzu u.U. ein größerer Anteil an spezifischer Varianz in Kauf genommen werden muss.

Da auch die Faktortreue ein minderungskorrigierter Validitätskoeffizient ist, stellt die Komposition von gepufferten Skalen im Sinne von Cattell letztlich ein Problem der Validitätsmaximierung dar. Wenn man Gewichtungen der Items zulässt, dann kann man die Validität eines Test maximieren, indem man mithilfe der multiplen Regression Gewichte für die einzelnen Variablen bestimmt (Lord & Novick, 1968). Die Schätzung der Regressionsgewichte über multiple Regression ist sowohl bei Vorliegen eines externen Kriteriums als auch im Rahmen der Faktorenanalyse (Thurstone, 1937) möglich. Der multiple Korrelationskoeffizient gibt dann die Validität des gewichteten Summenscores an. Da die multiple Korrelation durch die Entfernung von Prädiktoren nicht ansteigen kann, kann die Elimination von Items hier nur eine

ökonomischere Erfassung des Merkmals zum Ziel haben[6]. Hocking (1976) beschreibt Ansätze, die bei vorgegebner Anzahl an Prädiktoren die multiple Korrelation mit einem Kriterium maximieren.

Bei ungewichteten Items kann die Entfernung einzelner Items dagegen die Validität erhöhen (vgl. Ulrich, 1985; Burisch, 1997). Allerdings können bei ungewichteten Items (in der Stichprobe) grundsätzlich keine höheren Validitäten erreicht werden als bei (optimal) gewichteten Items. Der Verzicht auf eine Gewichtung ist nämlich äquivalent dazu, dass man nur eine Gewichtung mit 1 (Item ist in Skala enthalten) und 0 (Item wird entfernt) zulässt. Da die Regressionsgewichte BLUE (best linear unbiased estimators) sind, kann man jedoch nicht erwarten, dass ungewichtete Items in der zugrundeliegenden Population höher mit dem Kriterium korrelieren (Bryson, 1972).

Mit der Selektion ungewichteter Items zur Maximierung der Validität haben sich eine Reihe von Autoren beschäftigt (z.B. Horst, 1936, Gulliksen, 1950, Gleser & DuBois, 1951, Green, 1954, Webster, 1956, Darlington & Bishop, 1956). Diese Autoren versuchen, Kennwerte zu entwickeln, welche die Nützlichkeit von (ungewichteten) Items zur Steigerung der Validität erfassen sollen. Diese Kennwerte sind allesamt Funktionen der Itemvalidität, der Itemtrennschärfe sowie der Streuungen von Item und Test. Die Testvalidität lässt sich zwar als Funktion der Streuungen, Trennschärfen und Validitäten der einzelnen Items ausdrücken (vgl. Krauth, 1995, S. 277):

$$\begin{aligned}\rho_{T,C} &= \sum_{i=1}^{N}\sigma_i\rho_{i,C} \Big/ \sum_{i=1}^{N}\sigma_i\rho_{i,T} \\ &= \left(\frac{1}{N}\sum_{i=1}^{N}\sigma_i\rho_{i,C}\right) \Big/ \left(\frac{1}{N}\sum_{i=1}^{N}\sigma_i\rho_{i,T}\right)\end{aligned}$$

[2.2-1]

Jede Veränderung des Tests führt aber zwangsläufig zu veränderten Itemtrennschärfen. Die von Gulliksen (1950) empfohlene Methode zur Abschätzung der Itemvalidität anhand von [2.2-1] überschätzt daher die Validität eines verkürzten Tests. Der Ausmaß des Schätzfehlers wird umso größer, je stärker der entfernte Teil des Tests mit dem Rumpftest korreliert und je größer die Standardabweichung des entfernten Teils im Verhältnis zum Rumpftest ist (Green, 1954).

---

[6] Der Ausschluss von Variablen, die (in der Population) *gar keinen* Beitrag zur multiplen Korrelation leisten, verringert jedoch auch den Schätzfehler der Regressionsgewichte (Hocking, 1976).

Auch die anderen vorgeschlagenen Kennwerte für die Nützlichkeit von Items stellen nicht in Rechung, dass sich der Test während der Itemselektion ständig verändert. Die Nützlichkeit eines Items zur Sicherung der Validität hängt jedoch grundsätzlich davon ab, welche Items sonst noch in den Test aufgenommen werden. Daher erfordert die Maximierung der Validität die Berücksichtigung der gesamten Varianz-Kovarianzmatrix von Items und Kriterium (Gulliksen, 1950). Dann lässt sich die Validität jedes möglichen Subtests genau bestimmen (vgl. Ulrich, 1985):

$$\rho\left(\sum_{i=1}^{k} X_i, C\right) = \frac{\sigma\left(\sum_{i=1}^{k} X_i, C\right)}{\sigma\left(\sum_{i=1}^{k} X_i\right)\sigma(C)}$$

$$= \frac{\sum_{i=1}^{k}\sigma(X_i, C)}{\sigma(C)\cdot\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma(X_i, X_j)}}$$

$$= \frac{\sum_{i=1}^{k}\rho(X_i, C)\sigma(X_i)}{\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k}\rho(X_i, X_j)\sigma(X_i)\sigma(X_j)}}$$

[2.2-2]

Die exakte Berechnung der Validität der zur Auswahl stehenden Subtests anhand dieser Formel scheiterte früher jedoch an der begrenzten Rechnerkapazität.

Seit Aufkommen der computerunterstützten Datenauswertung stellt die Berücksichtigung sämtlicher Itemkorrelationen bei der Berechnung der Validität eines einzelnen Subtests kein Problem mehr dar. Die Validität eines Summenscores lässt sich bei bekannten Itemvaliditäten und Iteminterkorrelationen in Sekundenbruchteilen berechnen. Dennoch ist die Berechnung der Validität aller möglichen Subtests auch bei der heutigen Rechnerleistung nur bis zu Itempools aus etwa n=20 Items in vertretbarer Zeit zu bewältigen, da die Validität von 2^n-1 Tests zu berechnen ist. Da die möglichen Kombinationen exponentiell mit der Anzahl der Items ansteigen, ist nicht absehbar, ob es je gelingen wird, auch für Itempools aus 100 oder mehr Items die Validität aller Kombination zu berechnen. Für die Variablenselektion im Rahmen der multiplen Regression ist es gelungen, Algorithmen zu entwickeln, die sicherstellen, dass jeweils die bestmöglichen Tests einer bestimmten Länge zusammengestellt werden, ohne dass hierzu die Validitäten aller möglichen Tests dieser Länge berechnet werden müssen (z.B. Furnival & Wilson, 1974). Diese Verfahren machen sich zunutze, dass die multiple Korrelation durch die

Hinzunahme weiterer Variablen nicht sinken kann. Dies lässt sich jedoch nicht auf ungewichtete Items übertragen, da hier die Validität durch weitere Items durchaus absinken kann (vgl. Ulrich, 1985; Burisch, 1997).

Krauth (1995) schlägt daher vor, per Zufall so viele Tests einer gewünschten Testlänge auszuwählen, wie die verfügbare Rechenzeit zulässt. Das von Krauth (1995) vorgeschlagene Verfahren hat jedoch einen grundlegenden Nachteil: Es erhält keinen Algorithmus zur Optimierung der Validität, sondern versucht lediglich, durch wahlloses Ausprobieren möglichst gute Itemkombinationen zu finden.

Das von Ulrich (1985; Toops, 1941) vorgeschlagene Verfahren (MAXVAL) hat dagegen wesentlich bessere Chancen, die beste oder zumindest sehr gute Itemkombinationen zu entdecken. Dazu werden schrittweise diejenigen Items in den Test aufgenommen, bei denen sich die höchste Validität ergibt (Vorwärts-Selektion). Im ersten Schritt wird immer das Item mit der höchsten Validität ausgewählt. Im zweiten und den darauf folgenden Schritten wird aus den verbleibenden Items jedoch nicht unbedingt das valideste ausgewählt, sondern dasjenige, bei dessen Aufnahme der resultierende Test die höchste Validität hat. Nachdem alle Items hinzugefügt wurden, werden schrittweise diejenigen Items entfernt, bei denen der verbleibende Test jeweils die höchst mögliche Validität hat (Rückwärts-Selektion). Nach der Entfernung eines Items wird jeweils überprüft, ob der resultierende Test valider ist als der Test entsprechender Länge aus der Vorwärts-Selektion. Bei n Items erfordert das MAXVAL-Verfahren nur die Berechnung der Validität von $n(n+1)$ Teiltests. Aber dieses Verfahren führt nur dann mit Sicherheit zu optimalen Lösungen, wenn die besten Itemkombinationen jeder Testlänge all diejenigen Items enthalten, die in den besten Itemkombinationen kürzerer Testlänge enthalten sind (vgl. Berk, 1978). In diesem Fall ergeben sich keine Unterschiede zwischen der Vorwärts- und der Rückwärtsselektion. Unterschiedliche Ergebnisse bei der Vorwärts- und Rückwärtsselektion sind jedoch weder eine hinreichende noch eine notwendige Bedingung dafür, dass die Berücksichtigung aller Itemkombinationen bessere Resultate liefert als das MAXVAL-Verfahren. Es ist also denkbar, dass Vorwärts- und Rückwärtsselektion nicht für jede Testlänge übereinstimmen und das MAXVAL-Verfahren dennoch für jede Testlänge die optimale Itemkombination findet. Umgekehrt ist jedoch selbst die Übereinstimmung von Vorwärts- und Rückwärtsselektion keine Garantie dafür, dass das MAXVAL-Verfahren für jede Testlänge die optimale Lösung findet. Der Vergleich der Ergebnisse der Vorwärts- und Rückwärtsselektion liefert also keine verlässlichen Hinweise darauf, ob das MAXVAL-Verfahren nur suboptimale Ergebnisse geliefert hat.

Beim MAXVAL-Verfahren wird zwar gezielt nach besonders aussichtsreichen Itemkombinationen gesucht, aber es enthält dennoch keinen Algorithmus zur Optimierung. Es gibt jedoch eine einfache Möglichkeit, sowohl das MAXVAL-Verfahren als auch das Zufallsverfahren von Krauth (1995) zu optimieren. Jede beliebige Itemkombination lässt sich nämlich in die optimale Itemkombination der jeweiligen Testlänge transformieren, indem man wiederholt Items aus der suboptimalen Itemkombination durch Items der optimalen Itemkombination austauscht. Eine Möglichkeit, die Validität eines Tests gegebener Länge zu optimieren, besteht also darin, dass man alle möglichen Vertauschungen eines *einzelnen* (!) Items (des Itempools), das nicht im Test enthalten ist, mit einem anderen, welches nicht im Test enthalten ist, durchführt. Anschließend wird diejenige Vertauschung realisiert, bei der sich die größte Steigerung der Validität ergibt. Diesen Vorgang wiederholt man so oft, bis sich bei keiner der möglichen Vertauschungen von einzelnen Items eine Verbesserung der Validität ergibt. Dieses Verfahren stellt sicher, dass die gewählte Kombination zumindest ein lokales Maximum der Testvalidität ist, das durch die Vertauschung von *einzelnen* Items nicht weiter verbessert werden kann. Es ist aber durchaus möglich, dass sich durch den simultanen Austausch von mehreren Items eine Verbesserung der Testvalidität erreichen ließe. Bei größeren Itempools ist also denkbar, dass es neben dem absoluten Maximum der Testvalidität mehrere lokale Maxima gibt. Daher empfiehlt es sich, den Optimierungsvorgang mit verschiedenen Itemkombinationen als Startwert zu beginnen. Diese Startwerte lassen sich per Zufall wie bei dem Verfahren von Krauth bestimmen. Besonders vielversprechende Startwerte, die häufig bereits lokale oder gar absolute Maxima der Testvalidität sind, kann man wie bei dem MAXVAL-Verfahren durch die schrittweise Erweiterung oder Eliminierung von Items erhalten. Dabei wird immer die bestmögliche Erweiterung oder Verkürzung der Skala durchgeführt.

Im Programmpaket SAS ist dieses Verfahren zur Variablenauswahl im Rahmen der multiplen Regression realisiert worden. Allerdings wird hier darauf verzichtet, mehrere Startwerte einer bestimmten Testlänge vorzugeben. Stattdessen wird der Test – nachdem sich durch Vertauschungen keine Verbesserungen der Validität mehr erreichen lassen – um dasjenige Item erweitert, welches zur größten multiplen Korrelation führt. Hierbei besteht die Gefahr, dass der Algorithmus nur ein lokales Maximum findet. Es ist jedoch auch eine Variante dieses Verfahrens realisiert worden, die u.U. ein weiteres lokales Maximum entdeckt. Dazu wird nicht diejenige Vertauschung von Prädiktoren vorgenommen, die zur größten Steigerung der multiplen Korrelation führt, sondern diejenige, welche zur geringst möglichen Steigerung führt. Obwohl dieses Verfahren sehr viel mehr Subtests überprüft, ist es wenig wahrscheinlich, dass dieses

Verfahren zu besseren Lösungen führt, da es direkt an dem absoluten Maximum vorbei iterieren kann, obwohl es dessen multiple Korrelation berechnet hat.

Eine einfachere Version des schrittweisen Vorgehens empfehlen Gleser und Dubois (1951). Zunächst wird für jedes einzelne Item überprüft, ob die Entfernung zu einer Steigerung der Validität führt. Dann werden in einem Schritt alle Items entfernt, die sich mindernd auf die Validität des ursprünglichen Tests ausgewirkt haben. Bei diesem Verfahren wird der Rechenaufwand also extrem gering gehalten. Da jedoch mehrere Items auf einmal entfernt werden, ist nicht sichergestellt, dass sich die Validität mit jeder weiteren Iteration erhöht. Bei der Überprüfung der Wirkung auf die Validität wurde nämlich davon ausgegangen, dass jeweils nur ein Item dem Test hinzugefügt wird. Da die Minimierung des Rechenaufwands bei der heutigen Rechnerleistung nicht mehr die Bedeutung hat wie früher, muss das Verfahren von Gleser und Dubois als überholt betrachtet werden.

Auch das von Cattell und Tsujioka (1964) empfohlene Ausbalancieren der Faktor Ladungen (Semipartialkorrelationen oder standardisierten Regressionskoeffizienten) führt gerade bei korrelierten Faktoren nicht unbedingt zu optimalen Ergebnissen, da hierbei nicht in Rechnung gestellt wird, dass systematische Fehler bei korrelierten Faktoren zu gravierenderen Verzerrungen führen als bei unkorrelierten Faktoren. Daher sollte auch für die Maximierung der Faktorttreue sensu Cattell einer der oben beschriebenen, multivariaten Ansätze zur Optimierung der Validität verwendet werden. Dann sollte man jedoch die minderungskorrigierten Itemkorrelationen zugrunde legen.

2.2.3 Empirische Ergebnisse

Wie bereits erwähnt, sind die Ideen von Cattell kaum aufgegriffen worden, da eine Vielzahl von Studien belegt, dass der von ihm entwickelte 16 PF nicht nur hinsichtlich der klassischen Testgütekriterien schlecht abschneidet, sondern dass auch die von Cattell selbst entwickelten Kennwerte wie etwa die Faktortreue sehr niedrig sind. Da die Konstruktion des 16 PF jedoch weitgehend im Dunkeln liegt (Amelang & Bartussek, 2001), lassen diese Befunde kaum eine Bewertung der von Cattell und Tsujioka (1964) vorgeschlagenen Prinzipien der Skalenkonstruktion zu. Daher wird auf die umfangreiche Literatur über den 16 PF nicht näher eingegangen. Der interessierte Leser sei auf die Besprechung von Cattells Werk in Amelang und Bartussek (2001) verwiesen.

Auch zu den anderen besprochenen Selektionsmethoden liegen kaum aussagekräftige Befunde vor. Burisch (1997) hat das MAXVAL-Verfahren in zwei Stichproben (N=138 bzw. N=158) auf

die Items des Freiburger Persönlichkeitsinventars (FPI; Fahrenberg, Selg & Hampel, 1973) angewendet. Bei Kreuzvalidierung der ausgewählten Skalen zeigten sich anfangs keine besonders hohe Validitäten. In einem zweiten Versuch wurden die Items zunächst den Skalen zugeordnet, für die sie, nach Meinung einer Stichprobe von 51 Personen, prototypisch sind. Dabei durften einzelne Items durchaus mehreren Skalen zugeordnet werden. Bei Anwendung des MAXVAL-Verfahren auf die nach den Prototypizitätseinschätzungen revidierten Skalen ergab sich bei einigen Skalen auch nach Kreuzvalidierung eine relativ hohe Validität. In der jeweiligen Analysestichprobe erreichte die Validitätscharakteristik (= Zusammenhang von Testlänge und Validität) ihr Maximum bei einer Testlänge von zwischen 3 und 13 Items. Die Rangfolge mit der die Items in die Skalen aufgenommen wurden, unterschied sich jedoch in beiden Stichproben. Daher wurden nur solche Items in die Endform der Skalen aufgenommen, die in beiden Stichproben in den Test aufgenommen wurden, bevor die Validitätscharakteristik ihr Maximum erreicht hatte. Diese sogenannten Minimax-Skalen enthielten zwei bis drei Items und waren in einer dritten Stichprobe genauso valide wie die Originalskalen des FPI. Es wurde jedoch nicht mitgeteilt, welche Validität die vollständigen anhand der Prototypizitätsratings revidierten Skalen in dieser Stichprobe erreichten.

Auch wenn die Studie von Burisch (1997) ein Hinweis für die Nützlichkeit des MAXVAL-Verfahrens sein mag, ist das methodische Vorgehen nicht sehr überzeugend. So ergaben sich erst beim zweiten Versuch hohe Validitäten und auch hier nur bei einigen Skalen. Schwerer wiegt jedoch, dass sich theoretisch schwer begründen lässt, dass die Reorganisation anhand der Prototypizität zu einer besseren Performanz des MAXVAL-Verfahrens führt. Eine hohe Prototypizität mag zwar eine hohe Inhaltsvalidität der Items sichern und sich auf diesem Weg günstig auf die Validität der Skalen auswirken. Das MAXVAL-Verfahren zielt aber gar nicht darauf ab, inhaltlich homogene Skalen zusammenzustellen. Gerade bei heterogenen Itempools sollte das MAXVAL-Verfahren überlegen sein, da es über die Ausnutzung von Suppressionseffekten kriteriumsirrelevante Varianz unterdrücken kann (Ulrich, 1985). Da die Itemselektion beim MAXVAL-Verfahren nicht auf Itemkennwerten beruht, sondern auf Skalenkennwerten, und auf die Auswahl von besonders guten Itemkombinationen abzielt, lässt sich auch die Bildung der Minimax-Skalen nur schwer begründen. Sie widerspricht der Logik des MAXVAL-Verfahrens. Wenn beispielsweise zwei Items denselben kriteriumsirrelevanten Varianzanteil eines dritten Items nivellieren können, so kann es durchaus sein, dass in zwei verschiedenen Stichproben nur eines der beiden Suppressoritems vom MAXVAL-Verfahren aufgenommen wird. Bei der Bildung der Minimax-Skalen würden die beiden Suppressoritems daher u.U. nicht berücksichtigt werden. Die wenig befriedigenden Ergebnisse des MAXVAL-

Verfahrens in der Studie von Burisch (1997) könnten jedoch auch darauf zurückzuführen sein, dass erst bei sehr großen Stichproben die Regression zur Mitte, die in der Studie von Burisch zu beobachten war, die Vorteile der Ausnutzung von Suppressionseffekten nicht mehr zunichte macht.

2.2.4 Methodische Probleme

Bisher wurde bei der Diskussion der verschiedenen Verfahren zur Selektion von Subtests implizit davon ausgegangen, dass die Kovarianzmatrix der Items bekannt ist. In empirischen Anwendungen lassen sich jedoch nur fehlerbehaftete Schätzungen der Itemkovarianzen ermitteln. Die Itemkombination, die in einer vorliegenden Stichprobe die höchste Validität hat, ist nicht notwendigerweise auch die beste Itemkombination in der Population. Aufgrund der Regression zur Mitte ist dieser Fall sogar sehr unwahrscheinlich, wenn eine Vielzahl von Subtests verglichen wird. Stichprobenfehler bei der Variablenselektion sind ein schwieriges statistisches Problem. Im Rahmen der multiplen Regression sind bisher noch keine adäquaten Lösungen vorgeschlagen worden (vgl. Hocking, 1976; Werner, 1997). Auch hier ist mit überhöhten Schätzungen des multiplen Korrelationskoeffizienten zu rechnen, wenn Prädiktoren anhand von Stichprobendaten ausgewählt werden. Mit Monte-Carlo Simulationen lassen sich zwar Signifikanztests für unkorrelierte Prädiktoren entwickeln, die bei korrelierten Prädiktoren zu konservativen Entscheidungen führen (Diehr & Hoflin, 1974; Wilkinson, 1979). Diese Signifikanztests überprüfen jedoch nur, ob die multiple Korrelation der (ausgewählten) Prädiktoren mit dem Kriterium von null verschieden ist. Sie geben keine Auskunft darüber, in welchem Ausmaß der multiple Korrelationskoeffizient in der Stichprobe die wahre multiple Korrelation überschätzt, wenn die Nullhypothese nicht gilt. Man darf gespannt sein, ob diese Frage jemals geklärt wird, da „the distribution theory of the sample R^2 in best subset regression is hopelessly complex“ (Diehr & Hoflin, 1974, S. 317).

Bei ungewichteten Items sind Stichprobenfehler jedoch eher einer theoretischen Analyse zugänglich. Die Validitätskoeffizienten der einzelnen Subtests sind nämlich einfache Produkt-Moment Korrelationen. Wenn sowohl der Subtest als auch das Kriterium annähernd normalverteilt sind, dann ist der Fehler der Fisher-Z transformierten Validitätskoeffizienten, annähernd normalverteilt mit Varianz $\sigma_\varepsilon^2 = (N-3)^{-1}$, wobei N der Umfang der Personenstichprobe ist (Bortz, 1999). Die Schätzungen der wahren Validitätskoeffizienten durch Stichprobenkorrelationen lassen sich als fehlerbehaftete Messwerte im Sinne der klassischen Testtheorie auffassen (vgl. Kapitel 1.1). Wenn die gemeinsame Verteilung der Fisher-Z transformierten Validitätskoeffizienten ebenfalls annähernd normal ist, dann lässt sich eine

Punktschätzung für die wahren Validitätskoeffizienten angeben, welche die Regression zur Mitte berücksichtigt (vgl. Müller & Moosbrugger, 1985):

$$\hat{\rho}_{T,C} = \left(1 - \frac{\sigma_\varepsilon^2}{s_V^2}\right) \cdot r_{T,C} + \frac{\sigma_\varepsilon^2}{s_V^2} \cdot \bar{V}$$

[2.2-3]

$\bar{V}$ und s_V^2 sind Mittelwert und Varianz aller Fisher-Z transformierten Validitätskoeffizienten in der Stichprobe. Die Punktschätzung des wahren Validitätskoeffizienten ist also eine gewichtete Summe aus der Validität des betreffenden Subtests und dem Mittelwert der Validität aller verglichenen Subtests, wobei die Validität mit der Reliabilität der geschätzten Validitäten und der Mittelwert aller Subtests mit der Differenz der Reliabilität zu eins gewichtet wird (vgl. Müller & Moosbrugger, 1985).

Interessanter als eine Punktschätzung des wahren Validitätskoeffizienten ist jedoch die Frage, ob der Subtest mit der maximalen Validität in der Stichprobe auch in der Population eine vergleichsweise hohe Validität hat. Dies lässt sich durch eine τ-Normierung der Validitätskoeffizienten untersuchen (vgl. Müller & Moosbrugger, 1985). Für den wahren z-Wert $\zeta_{T,C}$ des Validitätskoeffizienten ergibt sich demnach folgende Punktschätzung:

$$\hat{\zeta}_{T,C} = \sqrt{1 - \frac{\sigma_\varepsilon^2}{s_V^2}} \cdot \frac{r_{T,C} - \bar{V}}{s_V}$$

[2.2-4]

Dieser Wert lässt sich auch in einen Normwert auf einer Prozentrangskala transformieren.

Bei den bisherigen Überlegungen wurde allerdings nicht berücksichtigt, dass die Population der Subtests finit ist. Da hier jedoch nur die Werte der besten Subtests interessieren und meist nur ein kleiner Teil der Subtests untersucht wird, sind dadurch keine größeren Verzerrungen zu erwarten. Schwerer wiegt die Tatsache, dass die Voraussetzungen für die verwendeten Formeln ([2.2-3], [2.2-4]) nicht gegeben sind, da die Fehler bei der Schätzung von Validitätskoeffizienten verschiedener Subtests nicht unabhängig sind. Dies liegt daran, dass jedes einzelne Item in einer Vielzahl von Subtests enthalten ist. Dies führt dazu, dass die Fehler (bei der Schätzung) der Validitätskoeffizienten der verschiedenen Subtests positiv korreliert sind. Folglich ist die Regression zur Mitte geringer, als bei unkorrelierten Fehlern zu erwarten wäre. Wenn die Korrelation der Fehler eins wäre, gäbe es gar keine Regression zur Mitte hin, da sämtliche Validitätskoeffizienten dann lediglich um eine Konstante verschoben wären. Die Varianz der

Validitätskoeffizienten in der Stichprobe entspräche in diesem Fall genau der Varianz der Validitätskoeffizienten in der Population. Schätzfehler der Validität von Tests, welche viele gemeinsame Items enthalten, dürften tendenziell in dieselbe Richtung gehen (d.h. positiv korreliert sein). Dies führt tendenziell zu konservativen Schätzungen der (τ-normierten) Validität des besten Subtests in der Stichprobe. Schätzungen der Validität der ausgewählten Subtests anhand von [2.2-3] und [2.2-4] haben demnach einen Bias zur Mitte hin.

Die Größe der Korrelationen zwischen den Schätzfehlern der Validitätskoeffizienten der einzelnen Subtests ist also von entscheidender Bedeutung für die Größe der Fehler bei der Itemselektion anhand von Stichprobendaten. Je höher der Anteil der gemeinsamen Items, desto größer sind die Korrelationen der Schätzfehler bei der Bestimmung der Validität. Wie viele Items haben zwei Tests der Länge k aus einem Itempool aus n Items im Durchschnitt gemeinsam?

Ein Test der Länge k lässt sich in jeden anderen Test gleicher Länge transformieren, indem man diejenigen Items, die nicht in dem anderen Test enthalten sind, durch solche austauscht, die nur in dem anderen Test enthalten sind. Wenn m Vertauschungen notwendig sind, dann ist bei beiden Tests der Anteil derjenigen Items, die auch in dem anderen Test enthalten sind gleich $(k-m)/k$. Es können bis zu k Items ausgetauscht werden, wenn der Test nicht mehr als die Hälfte aller Items enthält; ansonsten können maximal $n-k$ Items ausgetauscht werden. Es gibt $\binom{k}{m}$ Möglichkeiten, um m Items zu entfernen. Zu jeder dieser Möglichkeiten gibt es $\binom{n-k}{m}$ verschiedene Möglichkeiten, um die entfernten Items zu ersetzen. Insgesamt gibt es $\binom{n}{k}$ Tests mit der Länge k. Daraus ergibt sich folgende Formel für den durchschnittlichen Anteil gemeinsamer Items bei Tests mit gleicher Testlänge k:

$$\frac{\sum_{m=1}^{\min(k,n-k)} \binom{k}{m}\binom{n-k}{m}\frac{k-m}{k}}{\binom{n}{k}-1} = \frac{\frac{k}{n}\binom{n}{k}-1}{\binom{n}{k}-1} = \frac{\frac{k}{n}-\binom{n}{k}^{-1}}{1-\binom{n}{k}^{-1}}$$

[2.2-5]

Die durchschnittliche Anzahl gemeinsamer Items zweier Tests nimmt demnach annähernd linear mit der Testlänge k zu.

Die hier vorgestellte Ableitung liefert eine Erklärung für die von Berk (1978) angestellte Beobachtung, dass die multiple Korrelation bei Anwendung von Verfahren zur Variablenselektion besonders stark überschätzt wird, wenn nur ein geringer Teil der Prädiktoren ausgewählt wird. Es kann daher davon ausgegangen werden, dass auch bei Anwendung des MAXVAL-Verfahrens auf ungewichtete Items die Regression zur Mitte mit zunehmender Skalenlänge abnimmt.

Der Stichprobenfehler bei der Schätzung der Kovarianzmatrix betrifft zwar prinzipiell alle Selektionsmethoden, aber je größer die Anzahl der Subtests ist, die ein Verfahren berücksichtigt desto mehr dürfte die Validität der selegierten Subtests überschätzt werden. Da Verfahren, die auf Kennwerten von Skalen beruhen, in der Regel eine Vielzahl möglicher Itemkombinationen untersuchen, dürfte die Regression zur Mitte hier deutlich ausgeprägter sein. Da Stichprobenfehler mit zunehmendem Umfang der Personenstichprobe abnehmen, sind erst bei hinreichend großen Stichproben Vorteile durch die aufwendigeren, multivariaten Verfahren zu erwarten. Neben dem Umfang der Personenstichprobe dürften die Eigenschaften des Itempools von Bedeutung für die Effektivität der verschiedenen Selektionsverfahren sein.

Die Determinanten der Effektivität der verschiedenen Strategien der Itemselektion sind Gegenstand des empirischen Teils dieser Arbeit (Kapitel 5 und 6). Dabei werden die wichtigsten der hier besprochenen Ansätze zur Itemselektion anhand empirischer sowie simulierter Datensätze verglichen. Neben dem Vergleich von Methoden, die auf Validitätsdaten basieren, mit Ansätzen, die eine Homogenisierung anstreben, wird dabei der Frage nachgegangen, ob die in diesem Abschnitt beschriebenen multivariaten Methoden der Itemselektion den zuvor besprochenen, herkömmlichen Methoden der Skalenkonstruktion überlegen sind. Multivariate Methoden der Skalenkonstruktion versuchen, sich Suppressionseffekte zunutzte zu machen. Bisher wurden eher die inhaltlichen Grundlagen und das praktische Vorgehen im Rahmen dieses Ansatzes erörtert. Eine exakte Definition und Beschreibung des Suppressionskonzeptes steht bisher noch aus. Der folgende Abschnitt ist daher ganz diesem Thema gewidmet.

3 Suppression

In diesem Kapitel soll als Fortentwicklung des Ansatzes von Cattell und Tsujioka (1964) eine mathematische Definition des Suppressionsbegriffs in der Testtheorie vorgeschlagen werden. Dazu werden die einzelnen Items durch einfache, lineare Regression in eine Komponente zerlegt, die linear zu den wahren Werten des Kriteriums ist, und in ein Residuum, dass mit den wahren Werten des Kriteriums nicht korreliert. Wenn die Residualanteile der Items überwiegend negativ korrelieren, so hat die kriteriumsirrelevante Varianz einen deutlich geringeren Anteil an der Gesamtvarianz des Tests als wenn die Residualanteile unkorreliert sind. Daher ist es sinnvoll in diesem Fall von Suppression zu sprechen. Das Gegenteil von Suppression ist Redundanz. Sie liegt vor, falls die Residualanteile dagegen überwiegend positiv korrelieren. Dann kumuliert die kriteriumsirrelevante Varianz deutlich stärker als bei unkorrelierten Residualanteilen. Bevor die hier skizzierte Definition des Suppressionsbegriffs in Kapitel 3.2 mathematisch präzisiert wird, soll im nächsten Abschnitt zunächst die Suppressionskonzepte eingegangen werden, die im Rahmen der multiplen Regression entwickelt wurden.

3.1 Suppression in der multiplen Regression

Die im vorangegangenen Kapitel beschriebenen Methoden zur Maximierung der Validität durch die Itemselektion anhand von Skalenkennwerten (Kapitel 2.2) lassen sich allesamt als Versuch verstehen, gepufferte Skalen im Sinne von Cattell und Tsujioka (1964) zu entwickeln. Die Pufferung von Skalen ist nach Cattell und Tsujioka (1964) die Folge von Suppressionseffekten, d.h. von der Unterdrückung von Varianzanteilen, die nichts zur Vorhersage des Kriteriums beitragen.

Cattell und Tsujioka entwickelten ihr Konzept von Suppression ohne dabei Bezug auf die bereits vorhandene Literatur über Suppressoren im Rahmen der multiplen Regression zu nehmen (z.B. Horst, 1941; Lubin, 1957). Horst (1941) hat den Suppressionsbegriff in die statistische Psychologie eingeführt. Unter einem Suppressor versteht Horst einen Prädiktor der nicht mit dem Kriterium korreliert, aber aufgrund von Korrelationen mit anderen Prädiktoren dennoch zur Varianzaufklärung beiträgt. Der Suppressionsbegriff im Rahmen der multiplen Regression wurde allerdings seither weiter entwickelt (Conger, 1974; Velicer, 1978; Holling, 1981). Laut Conger (1974) liegt Suppression genau dann vor, wenn das standardisierte Regressionsgewicht eines Prädiktors (bei identischem Vorzeichen) größer ist als die bivariate Kriteriumskorrelation. Die von Horst beschriebene Konstellation ist demnach lediglich ein spezieller Fall von Suppression. Velicer (1978) schlägt statt dessen vor, nur dann von Suppression zu sprechen, wenn die quadrierte Semipartialkorrelation eines Prädiktors größer ist als dessen bivariater

Determinationskoeffizient. Holling (1981) ist der Auffassung, dass man den Suppressionsbegriff von Velicer auf solche Fälle begrenzen sollte, in denen die Semipartialkorrelation und die bivariate Korrelation eines Prädiktors dasselbe Vorzeichen haben.

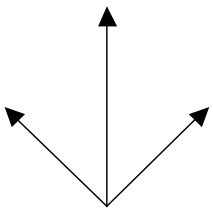
Im folgenden Kapitel soll gezeigt werden, warum der Suppressionsbegriff aus der multiplen Regression für die Definition von Suppression in der Testtheorie nicht geeignet ist. Außerdem soll das von Cattell und Tsujioka für die Testtheorie entwickelte Suppressionskonzept mathematisch präzisiert werden.

3.2 Suppression in der Testkonstruktion

3.2.1 Vergleich von Suppression in Testkonstruktion und der multiplen Regression

Als paradigmatisches Beispiel für Suppressionseffekte diskutieren Cattell und Tsujioka (1964) den Fall von zwei orthogonalen (standardisierten) Prädiktoren, deren Summe mit dem zu messenden Persönlichkeitsfaktor zu eins korreliert. Die perfekte Vorhersage des Kriteriums ist in diesem Fall darauf zurückzuführen, dass die Ladungen auf dem irrelevanten Persönlichkeitsfaktor ausbalanciert werden.

Abbildung 1: Suppression sensu Cattell und Tsujioka (1964)



Bei der multiplen Regression werden die Prädiktoren jedoch immer so gewichtet, dass irrelevante Varianzanteile unterdrückt werden. Der Cattellsche Suppressionsbegriff würde demnach auf alle Prädiktoren im Rahmen der multiplen Regression zutreffen. Jeder der Prädiktoren, die an der multiplen Regression beteiligt sind, lässt sich nämlich additiv in zwei orthogonale Komponenten (Zufallsvariable) zerlegen, von denen eine mit den durch die multiple Regression vorhergesagten Werten zu eins korreliert, während die andere mit dem Kriterium nicht korreliert. In der Regressionsgleichung werden die Prädiktoren so gewichtet, dass sich die letzteren Komponenten (welche mit dem Kriterium nicht korrelieren) gegenseitig aufheben. Im Rahmen der multiplen Regression spricht man daher nur dann von Suppression, wenn die Unterdrückung von störenden Varianzanteilen so groß ist, dass der Beitrag eines Prädiktors zur

Varianzaufklärung größer ist als aufgrund seiner bivariaten Korrelation mit dem Kriterium zu erwarten wäre. Der Beitrag eines Prädiktors zur Varianzaufklärung lässt sich als standardisiertes Regressionsgewicht (Ansatz von Conger, 1974) oder als Semipartialkorrelation (Ansatz von Velicer, 1978 und Holling, 1981) quantifizieren.

Bei beiden Definitionen von Suppression bezieht sich der (implizite) Erwartungshorizont auf orthogonale Prädiktoren. Bei orthogonalen Prädiktoren entsprechen die standardisierten Regressionsgewichte sowie die Semipartialkorrelationen den bivariaten Kriteriumskorrelationen und der multiple Determinationskoeffizient ist die Summe der einfachen Determinationskoeffizienten.

3.2.2 Mathematische Definition von Suppression in der Testtheorie

Im Rahmen der klassischen Testtheorie ist es sicherlich unangemessen, Orthogonalität als Referenzpunkt zu wählen, da die Items eines Tests Indikatoren derselben Persönlichkeitseigenschaft sein sollen. Es sind also positiv korrelierte Variablen zu erwarten, deren Korrelationen mit externen Kriterien jeweils alle dasselbe Vorzeichen haben müssten. Sie wären also eher als redundante Prädiktoren zu klassifizieren. Im Folgenden soll daher eine mathematische Definition des Suppressorkonzepts entwickelt werden, welches für die Testtheorie angemessener ist.

Klassische, eindimensionale Testmodelle gehen davon aus, dass die Itemscores X_1, ... , X_N einer Skala T sich additiv aus einem wahrem Wert sowie einer unsystematischen Fehlervariable zusammensetzten (vgl. [1.1-2] auf S. 3), wobei die wahren Werte der Items jeweils zu eins korrelieren (τ-Kongenerität, vgl. [1.1-17] auf S. 10), während die Fehler unkorreliert sind. Dieses Modell ist ein sinnvoller Vergleichsmaßstab für die Definition eines testtheoretischen Suppressionsbegriffs.

Die Werte jedes Items X_i lassen sich in jedem Fall durch einfache, lineare Regression additiv in eine Komponente zerlegen, die linear zu den *wahren* Werten des Kriteriums ist, und in ein Residuum ω_i , das nicht mit den wahren Werten des Kriteriums korreliert. Wenn die Fehler der Items nicht mit dem Fehler des Kriteriums korrelieren gilt (vgl. [1.2-1] auf S. 14 und [1.2-6] auf S. 18):

$$
\begin{aligned}
X_i &= \beta_{0i} + \beta_{1i} \cdot \tau(C) + \omega_i \\
&= \beta_{0i} + \frac{\rho_{i,\tau(C)}}{\sigma_{\tau(C)}} \cdot \sigma_i \cdot \tau(C) + \omega_i \\
&= \beta_{0i} + \frac{\rho_{i,C}}{\sigma_C \cdot \mathrm{Rel(C)}} \cdot \sigma_i \cdot \tau(C) + \omega_i
\end{aligned}
$$

[3.2-1]

Das Residuum ω_i entspricht nur dann der Fehlerkomponente des Items, wenn das Kriterium und das Item τ-kongenerisch (vgl. [1.1-17] auf S. 10) und die Messfehler unkorreliert sind. Aber auch wenn das Item und das Kriterium nicht dieselbe latente Variable erfassen, lässt sich die Varianz des Items folgendermaßen in zwei Komponenten zerlegen:

$$
\begin{aligned}
\sigma_i^2 &= \rho_{i,\tau(C)}^2 \sigma_i^2 + \sigma^2(\omega_i) \\
&= \rho_{i,\tau(C)}^2 \sigma_i^2 + \left(1 - \rho_{i,\tau(C)}^2\right)\sigma_i^2 \quad , \\
&= \frac{\rho_{i,C}^2}{\mathrm{Rel(C)}} \sigma_i^2 + \left(1 - \frac{\rho_{i,C}^2}{\mathrm{Rel(C)}}\right)\sigma_i^2
\end{aligned}
$$

[3.2-2]

wobei der erste Summand den Varianzanteil repräsentiert, den das Item mit dem Kriterium teilt und der zweite Summand die davon unabhängige Residualvarianz. Nur wenn alle Items des Tests $T = X_1 + \ldots + X_N$ hinsichtlich des Kriteriums τ-kongenerisch sind und zudem die Messfehler der Items unkorreliert sind, müssen die Residualkomponenten der verschiedenen Items $\omega_1, \ldots, \omega_N$ unkorreliert sein. Andernfalls sind sowohl positive als auch negative Kovarianzen zwischen den Residualkomponenten möglich. Die mit dem Kriterium geteilten Komponenten der verschiedenen Items korrelieren in jedem Fall jeweils zu eins miteinander. Für die Varianz des Tests T ergibt sich demnach (vgl. [3.2-1] und [3.2-2]):

$$
\begin{aligned}
\sigma_T^2 &= \sigma^2\left(\sum_{i=1}^{N} X_i\right) \\
&= \sigma^2\left(\sum_{i=1}^{N} \beta_{0i} + \beta_{1i} \cdot \tau(C) + \omega_i\right) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} \sigma\left(\beta_{1i} \cdot \tau(C), \beta_{1j} \cdot \tau(C)\right) + \sum_{i=1}^{N}\sum_{j=1}^{N} \sigma\left(\omega_i . \omega_j\right) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} \beta_{1i} \cdot \beta_{1j} \cdot \sigma_C^2 \cdot \mathrm{Rel(C)} + \sum_{i=1}^{N} \sigma^2(\omega_i) + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \sigma\left(\omega_i . \omega_j\right) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\rho_{i,C}\rho_{j,C}}{\mathrm{Rel(C)}} \cdot \sigma_i \sigma_j + \sum_{i=1}^{N}\left(1 - \frac{\rho_{i,C}^2}{\mathrm{Rel(C)}}\right) \cdot \sigma_i^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \sigma\left(\omega_i, \omega_j\right)
\end{aligned}
$$

[3.2-3]

Der erste Summand gibt den Varianzanteil wieder, den der Test *T* mit dem Kriterium teilt. Je größer dieser Summand ist, desto höher ist die Kovarianz von Test und Kriterium. Die beiden anderen Summanden gehen auf die Residualkomponenten der Items und deren Kovarianzen zurück. Die Kovarianz des Tests mit dem Kriterium ist unabhängig von der Größe der letzten beiden Summanden in [3.2-3]. Da die beiden Summanden jedoch zur Varianz des Tests beitragen, mindern sie die Korrelation des Tests mit dem Kriterium. Wenn die Summe der Kovarianzen der Residualkomponenten negativ ist, kommt es zu einer Suppression der störenden Residualvarianz. Dann ist die Validität größer als bei Items mit entsprechender Validität, die sowohl untereinander als auch hinsichtlich des Kriteriums τ-kongenerisch sind (und deren Messfehler nicht korrelieren).

Wir definieren daher: Suppression zwischen den Items eines Tests liegt genau dann vor, wenn folgende Bedingung erfüllt ist (vgl. [3.2-3]):

$$2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sigma\left(\omega_i,\omega_j\right)=\sigma_T^2-\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\rho_{i,C}\rho_{j,C}}{\mathrm{Rel(C)}}\cdot\sigma_i\sigma_j-\sum_{i=1}^{N}\left(1-\frac{\rho_{i,C}^2}{\mathrm{Rel(C)}}\right)\cdot\sigma_i^2<0$$

[3.2-4]

$$\Leftrightarrow\qquad \sigma_T^2\cdot\mathrm{Rel(C)}<\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{i,C}\rho_{j,C}\cdot\sigma_i\sigma_j+\sum_{i=1}^{N}\left(\mathrm{Rel(C)}-\rho_{i,C}^2\right)\cdot\sigma_i^2$$

$$\Leftrightarrow\qquad \sigma_T^2\cdot\mathrm{Rel(C)}<\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{i,C}\rho_{j,C}\cdot\sigma_i\sigma_j-\sum_{i=1}^{N}\rho_{i,C}^2\cdot\sigma_i^2+\mathrm{Rel(C)}\cdot\sum_{i=1}^{N}\sigma_i^2$$

$$\Leftrightarrow\ \mathrm{Rel(C)}\left(\sigma_T^2-\sum_{i=1}^{N}\sigma_i^2\right)<\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{i,C}\rho_{j,C}\cdot\sigma_i\sigma_j-\sum_{i=1}^{N}\rho_{i,C}^2\cdot\sigma_i^2$$

[3.2-5]

$$\Leftrightarrow\qquad \begin{cases} \mathrm{Rel(C)}<\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{i,C}\rho_{j,C}\cdot\sigma_i\sigma_j-\sum_{i=1}^{N}\rho_{i,C}^2\cdot\sigma_i^2\right)\Big/\left(\sigma_T^2-\sum_{i=1}^{N}\sigma_i^2\right), & \text{falls } \left(\sigma_T^2-\sum_{i=1}^{N}\sigma_i^2\right)>0 \\ \mathrm{Rel(C)}>\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{i,C}\rho_{j,C}\cdot\sigma_i\sigma_j-\sum_{i=1}^{N}\rho_{i,C}^2\cdot\sigma_i^2\right)\Big/\left(\sigma_T^2-\sum_{i=1}^{N}\sigma_i^2\right), & \text{falls } \left(\sigma_T^2-\sum_{i=1}^{N}\sigma_i^2\right)<0 \\ \sum_{i=1}^{N}\rho_{i,C}^2\cdot\sigma_i^2<\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{i,C}\rho_{j,C}\cdot\sigma_i\sigma_j, & \text{falls } \left(\sigma_T^2-\sum_{i=1}^{N}\sigma_i^2\right)=0 \end{cases}$$

[3.2-6]

Da sich die Suppression in [3.2-3] bzw. [3.2-5] nur auf die Residualkomponenten ω_i der Items bezieht, sollte man in Abgrenzung zum Suppressionsbegriff in der multiplen Regression eher

von Fehlersuppression[7] sprechen. Wenn die Residualkomponenten der Items überwiegend positiv korreliert sind, könnte man von Fehlerredundanz sprechen. Sie ließe sich komplementär zur Fehlersuppression definieren, indem man „<“ in [3.2-5] durch „>“ ersetzt. Fehlerredundanz entspricht der Konstellation, die Cattell und Tsujioka (1964, S. 5) als „bloated specific disorder“ bezeichnet haben. Dass die Selektion von Items anhand ihrer Trennschärfe zwar zur Auswahl von validen Items, nicht jedoch zu validen Skalen führt (vgl. die Zusammenfassung der Ergebnisse von Yousfi, 2004b in Kapitel 2.1 auf S. 26), ist offensichtlich das Ergebnis von Fehlerredundanz, da bei der Selektion anhand der Trennschärfe auch die kriteriumsirrelevanten Varianzanteile kumulieren.

Auch sonst ist die hier vorgeschlagene Explikation des Suppressionsbegriffs mit dem Ansatz von Cattell und Tsujioka (1964) verträglich. So erfüllen unkorrelierte Items, deren Validitätskoeffizienten dasselbe Vorzeichen haben, immer die Bedingungen für Fehlersuppression ([3.2-5]). Dies lässt sich leicht nachvollziehen, wenn man bedenkt, dass die Kovarianz der Residualkomponenten das Produkt aus deren Standardabweichungen mit der Partialkorrelation der Items bei Herauspartialisierung der wahren Kriteriumswerte ist (vgl. [3.2-1] auf S. 43 und [3.2-2] auf S. 43).

$$\begin{aligned}\sigma\left(\omega_i,\omega_j\right) &= \rho\left(\omega_i,\omega_j\right)\cdot\sigma\left(\omega_i\right)\cdot\sigma\left(\omega_j\right)\\ &= \rho_{ij\bullet\tau_C}\cdot\sigma_i\cdot\sigma_j\cdot\sqrt{\left(1-\frac{\rho_{i,C}^2}{\mathrm{Rel}\left(C\right)}\right)\cdot\left(1-\frac{\rho_{j,C}^2}{\mathrm{Rel}\left(C\right)}\right)}\end{aligned}$$

[3.2-7]

Die Partialkorrelation von zwei unkorrelierten Variablen ist wiederum genau dann negativ, wenn deren Korrelationen mit der herauspartialisierten Variablen das gleiche Vorzeichen haben (vgl. die erste Zeile von [3.2-8] auf S. 46).

Bei einem Test, dessen Items τ-kongenerisch sind, ist Fehlersuppression dagegen unmöglich, falls die Messfehler der Items und des Kriteriums unkorreliert sind. Bei Anwendung der Verdünnungsformeln (vgl. [1.2-6] auf S. 18) auf die paarweisen Partialkorrelationen der Items

[7] Aber auch diese Bezeichnung könnte missverstanden werden, da die Suppression sich nicht auf die wechselseitige Unterdrückung unsystematischer Fehlervarianz im Sinne der klassischen Testtheorie, sondern vielmehr auf die Suppression systematischer kriteriumsirrelevanter Varianzanteile bezieht. Nur wenn man keine unkorrelierten Messfehler postuliert, können diese systematischen Varianzanteile nicht nur wahre, sondern auch Fehlervarianz im Sinne der klassischen Testtheorie sein.

bei Herauspartialisierung der wahren Kriteriumswerte ergibt sich nämlich (vgl. [1.1-17] auf S. 10):

$$\begin{aligned}\rho_{ij\bullet\tau_C} &= \frac{\rho\left(X_i, X_j\right) - \rho\left(X_i, \tau_C\right)\rho\left(X_j, \tau_C\right)}{\sqrt{1-\rho^2\left(X_i, \tau_C\right)}\sqrt{1-\rho^2\left(X_j, \tau_C\right)}} \\ &= \frac{\sqrt{REL_i \cdot REL_j} - \rho\left(\tau_i, \tau_C\right)\rho\left(\tau_j, \tau_C\right)\sqrt{REL_i \cdot REL_j}}{\sqrt{1-\rho^2\left(X_i, \tau_C\right)}\sqrt{1-\rho^2\left(X_j, \tau_C\right)}} \\ &= \frac{\sqrt{REL_i \cdot REL_j}\left(1-\rho\left(\tau_i, \tau_C\right)\rho\left(\tau_j, \tau_C\right)\right)}{\sqrt{1-\rho^2\left(X_i, \tau_C\right)}\sqrt{1-\rho^2\left(X_j, \tau_C\right)}} \geq 0\end{aligned}$$

[3.2-8]

Aus [3.2-5] (S. 44) und [3.2-8] geht unmittelbar hervor, dass (bei unkorrelierten Messfehlern der Items) immer Fehlerredundanz vorliegen muss, wenn die Items zwar untereinander, nicht jedoch mit dem Kriterium τ-kongenerisch sind. Dies unterstützt die Argumentation von Cattell und Tsujioka (1964), wonach sich bei Geltung klassischer Testmodelle die Anhäufung kriteriumsirrelevanter Varianz nur vermeiden lässt, wenn es gelingt, Items zu generieren, denen keinerlei systematische Fehler bei der Messung der intendierten Persönlichkeitseigenschaft anhaftet. Dies dürfte in der Praxis kaum zu erreichen sein.

3.2.3 Voraussetzungen für die praktische Anwendung

Das hier vorgestellte Kriterium für Suppression lässt sich immer dann empirisch überprüfen, wenn nicht nur die Itemvaliditäten, die Itemstreuungen sowie die Streuung des Tests, sondern auch die Reliabilität des Kriteriums anhand von Stichprobendaten geschätzt werden kann. Ob die Items eines Tests Suppressoren sind, lässt sich allerdings nicht allgemeingültig beantworten. Es kann durchaus sein, dass die Items hinsichtlich einer Kriteriumsvariable die Bedingungen für Suppression ([3.2-5] auf S. 44) erfüllen, während dies für ein anderes Kriterium nicht der Fall ist.

Bei der hier vorgestellten Explikation des Suppressionsbegriffs wurde lediglich vorausgesetzt, dass der Messfehler des Kriteriums nicht mit den Messfehlern der Items korreliert (durch die Verwendung von [1.2-6], S. 18 bei der Herleihung von [3.2-1], S. 43). Während die Annahme unkorrelierter Messfehler zwischen den Items eines Tests kaum zu rechtfertigen ist (wegen reaktionskontingenten Übungseffekten, Einflüssen von States bei Traittests, Ermüdung, Tendenzen zur konsistenten Selbstdarstellung), ist es aufgrund der unterschiedlichen Herkunft der Daten häufig durchaus angemessen zu postulieren, dass die Fehler (der Items) eines Tests nicht mit den Fehlern des Kriteriums korrelieren. Wenn die Fehler des Kriteriums mit den

Fehlern der Items dennoch korreliert sind, dann ist [3.2-4] (S. 44) keine sinnvolle Operationalisierung des Suppressionsbegriffs.

Korrelationen der Messfehler der Items sind dagegen durchaus mit der hier vorgestellten Definition von Suppression verträglich. Sie sind sogar eine mögliche Ursache für Fehlersuppression oder Fehlerredundanz. Bei τ-kongenerischen Items sind sie sogar die einzig mögliche Ursache für Fehlersuppression (nicht jedoch für Fehlerredundanz; vgl. [3.2-8] auf S. 46).

Das hier vorgestellte Suppressionskriterium dürfte demnach in den allermeisten Anwendungen eine sinnvolle Operationalisierung des Suppressionsbegriffs für Zwecke der Testtheorie und Testkonstruktion sein. Dies soll im empirischen Teil der Arbeit demonstriert werden.

4 Fragestellung der empirischen Studien

Im vorangegangenen Kapitel sind verschiedene Methoden der Skalenkonstruktion diskutiert worden. Im empirischen Teil der Arbeit sollen diese Methoden anhand verschiedener Datensätze verglichen werden. Dabei werden sowohl empirische als auch simulierte Datensätze herangezogen. Die Verwendung von simulierten Daten hat den Vorteil, dass die Populationsparameter (Validität, Trennschärfe etc.) bekannt sind. Dies ermöglicht eine genaue Analyse der statistischen Eigenschaften der aus den simulierten Daten gewonnenen Parameterschätzungen. Außerdem kann man so eine Vielzahl von Datensätzen erzeugen, bei denen verschiedene Merkmale der Population sowie der gezogenen Stichprobe gezielt variiert werden können. Auf diesem Wege lassen sich Aussagen darüber ableiten, von welchen Bedingungen die Effektivität der einzelnen Verfahren abhängt. Häufig werden die Ergebnisse derartiger Monte-Carlo Simulationen lediglich deskriptiv ausgewertet. Der experimentelle Ansatz dieser Studien ermöglicht jedoch durchaus die inferenzstatistische Überprüfung von Hypothesen, die sich auf Determinanten der Effektivität der verschiedenen Konstruktionsverfahren beziehen (Harwell, 1997). Dies stellt einen entscheidenden Vorteil gegenüber Studien an empirischen Datensätzen dar. Anhand empirischer Datensätze lassen sich diese Hypothesen nämlich kaum prüfen, da viele Merkmale der Personenstichprobe sowie des Itempools unbekannt sind. Anderseits ist die ökologische Validität von Aussagen, die aus simulierten Daten abgeleitet wurden, grundsätzlich fragwürdig. Nur wenn es gelingt, die Parameter in derartigen Studien so zu setzen, dass sie repräsentativ für empirische Personen- und Itempopulationen sind, ist zu erwarten, dass die Ergebnisse für empirische Anwendungen relevant sind.

4.1 Untersuchte Selektionsstrategien

In Kapitel 2 wurden verschiedene Ansätze zur Selektion von Items beschrieben. Wenngleich nicht jeder dieser Ansätze in der vorliegenden Arbeit realisiert werden konnte, sollte doch ein möglichst breiter Querschnitt von Verfahren untersucht werden. Insbesondere interessierte dabei die Frage, ob sich durch multivariate Methoden der Itemselektion, welche die gesamte Kovarianzmatrix der Items berücksichtigen, eine Verbesserung der Testgüte erreichen lässt. Außerdem sollte untersucht werden, inwieweit Selektionsstrategien, die nicht auf Validitätsdaten basieren, sondern eine Homogenisierung des Itempools anstreben, imstande sind, die Testvalidität zu sichern. Tabelle 1 gibt einen Überblick über die angewendeten Selektionsalgorithmen.

Tabelle 1: Einordnung der Selektionsalgorithmen

Ziel	*Univariat*	*Multivariat*
Validitätsmaximierung	• Itemvalidität	• MAXVAL • Optimierung der Validität • Vollständige Permutation
Homogenisierung	• Trennschärfe	• Optimierung der internen Konsistenz

Im einzelnen wurden die folgenden Selektionsalgorithmen untersucht:

- Zufällige Auswahl

  Die Items werden schrittweise in den Test aufgenommen, wobei die Reihenfolge per Zufall und ohne Berücksichtigung statistischer Kennwerte erfolgt. Nur solche Itemselektionsstrategien, die zu einer deutlich besseren Testgüte führen als die zufällige Auswahl, sollten bei der Testkonstruktion in Betracht gezogen werden. Eine wiederholte zufällige Auswahl zur Optimierung der internen Konsistenz oder der externen Validität, wie sie von Krauth (1995) empfohlen wird, wurde nicht untersucht, da dieses Verfahren wenig effizient sein dürfte. Da bei zufälliger Auswahl weder eine Homogenisierung der Skala noch eine Maximierung der Validität angestrebt wird, ist dieses Verfahren nicht in Tabelle 1 aufgenommen worden.

- Itemvalidität

  Die Items werden schrittweise anhand ihrer Validität in der Analysestichprobe in den Test aufgenommen.

- MAXVAL

  Dieses Verfahren wurde in Kapitel 2.2.2 auf S. 32 detailliert beschrieben. Die Items werden zunächst schrittweise in den Test aufgenommen (Vorwärtsselektion), wobei jeweils diejenigen Items ausgewählt werden, bei denen sich die höchste Validität in der Analysestichprobe ergibt. Nachdem alle Items in den Test aufgenommen wurden, werden schrittweise diejenigen Items aus dem Test entfernt (Rückwärtsselektion), bei deren Ausschluss sich die höchste Validität ergibt. Für jede Testlänge werden schließlich die Ergebnisse der Vorwärtsselektion und der Rückwärtsselektion verglichen und die bessere Alternative wird realisiert.

- Optimierung der Validität

  Dieses Verfahren wurde in Kapitel 2.2.2 auf S. 33 detailliert beschrieben. Es stellt eine Weiterentwicklung des MAXVAL-Algorithmus dar: Nach jeder Aufnahme eines Items bei der Vorwärtsselektion oder Entfernung eines Items bei der Rückwärtsselektion, wird solange wiederholt die beste aller möglichen Vertauschungen eines *einzelnen* im Test enthaltenen Items durch *ein* anderes nicht enthaltenes Item vorgenommen, bis sich durch den Austausch eines *einzelnen* Items keine Verbesserungen mehr erzielen lassen. Die in Kapitel 2.2.2 auf S. 33 skizzierte Erweiterung dieses Algorithmus um einen zufälligen Startpunkt wurde nicht realisiert, da die Rechenzeit in der Simulationsstudie sonst zu lang geworden wäre.

- Vollständige Permutation

  Da dieses Verfahren sehr rechenintensiv ist, konnte es nur bei den kleinsten Itempools der Simulationsstudie angewendet werden.

- Trennschärfe

  Die Items werden schrittweise anhand ihrer Trennschärfe in der Analysestichprobe in den Test aufgenommen. Als Trennschärfe wurde dabei jeweils die Produkt-Moment Korrelation des Itemscores mit der Summe aller Items des jeweiligen Itempools bestimmt. In Kapitel 2.1 wurde (auf S. 26) die Erwartung geäußert, dass die Bevorzugung trennscharfer Items sich ungünstig auf die Validität auswirkt.

- Optimierung der internen Konsistenz

  Dieses Verfahren funktioniert genauso wie die oben beschriebene Optimierung der Validität mit dem Unterschied, dass die Maximierung von Cronbachs α das Kriterium für die Itemauswahl in der Analysestichprobe ist. Da Cronbachs α erst ab einer Skalenlänge von zwei Items berechnet werden kann, wurde bei einer Skalenlänge von eins eine zufällige Auswahl der Items vorgenommen.

4.2 Hypothesen

Stichprobenumfang und Schätzfehler

Die Effektivität der verschiedenen Itemselektionsmethoden dürfte nicht zuletzt von der Größe der Personenstichprobe abhängen, anhand derer die statistischen Kennwerte berechnet werden, die zur Itemselektion herangezogen werden. Zwar ist bei allen Verfahren mit besseren Ergebnissen zu rechnen, wenn der Stichprobenumfang zunimmt, da die Messfehler dann kleiner sind. Allerdings dürften diejenigen Verfahren, die Skalenkennwerte analysieren (Kapitel 2.2.2

bzw. Spalte „Multivariat“ in Tabelle 1 auf S. 49), sehr viel empfindlicher auf Stichprobenfehler reagieren als Verfahren, die auf Itemkennwerten basieren (Kapitel 2.1 bzw. Spalte „Univariat“ in Tabelle 1 auf S. 49). Dies liegt daran, dass bei diesen Verfahren aus einer sehr großen Anzahl von Tests (Itemteilmengen) diejenige ausgewählt wird, die in der Stichprobe die besten Ergebnisse erzielt. Da hierbei sehr viel mehr Kennwerte verglichen werden als bei Verfahren, die nur Statistiken einzelner Items vergleichen, muss man davon ausgehen, dass auch die Regression zur Mitte sehr viel ausgeprägter ist (vgl. [2.2-3] auf S. 37 und [2.2-4] auf S. 37). Daher wurde die Größe der Personenstichprobe in den Simulationsstudien systematisch variiert.

Größe des Itempools

Aber auch die Größe des Itempools könnte sich auf die Effektivität der verschiedenen Verfahren auswirken. Die Menge möglicher Itemkombinationen nimmt exponentiell mit dem Umfang des Itempools zu. Vor allem bei umfangreichen Itempools dürfte es daher Itemkombinationen geben, die besser sind als diejenigen, die sich bei der Selektion anhand von Itemparametern ergeben. Allerdings ist auch das Problem der Regression zur Mitte in diesem Fall besonders gravierend (vgl. [2.2-3] auf S. 37 und [2.2-4] auf S. 37), so dass nur bei genauen Schätzungen der Populationsparameter anhand großer Personenstichproben Vorteile für multivariate Verfahren zu erwarten sind.

Anzahl der latenten Dimensionen

Es ist zu erwarten, dass der Nutzen der verschiedenen Selektionsstrategien auch von der dimensionalen Struktur des Itempools abhängt. Vorteile für Verfahren, die Itemkombinationen untersuchen, dürften sich eher bei solchen Itemmengen ergeben, denen mehr als eine Dimension zugrunde liegt, da man nur in diesem Fall markante Suppressionseffekte erwarten kann. Die Anzahl der zugrundeliegenden Dimensionen des Itempools kann selbstverständlich nur bei simulierten Daten experimentell variiert werden.

Testlänge

In Kapitel 2.1 wurde darauf hingewiesen, dass die Selektion anhand von Trennschärfen und interner Konsistenz meist wenig für die Sicherung der Validität bringt, wenn der Itempool nicht eindimensional ist. Die von Yousfi (2004b, siehe auch die Zusammenfassung auf S. 26) präsentierten Ableitungen beziehen sich jedoch auf die Elimination eines einzelnen Items. Welche Konsequenz die simultane Elimination von mehreren Items anhand dieser Kriterien hat, ist analytisch schwer zu beantworten. Dennoch erscheint es sinnvoll anzunehmen, dass die Selektion anhand der Trennschärfe oder der internen Konsistenz bei der Elimination weniger

Items kaum zu besseren Ergebnissen führt als die zufällige Auswahl. Wenn dagegen nur wenige trennscharfe Items eines validen Tests aufsummiert werden, kann man von der Trennschärfe deutlich bessere Ergebnisse erwarten als bei zufälliger Auswahl. Eine Maximierung der internen Konsistenz sollte sich dagegen allenfalls bei eindimensionalen Itempools günstig auf die Testvalidität auswirken.

Die Anzahl möglicher Tests ist genau dann maximal, wenn die Hälfte der Items im Test enthalten ist. Daher könnte man erwarten, dass multivariate Methoden, die ja die Vielfalt der möglichen Itemkombinationen gezielt nach besonders validen Tests durchforsten, besonders gut abschneiden, wenn etwa die Hälfte der Items in den Test aufgenommen wird. Aus [2.2-5] (S. 38) ergibt sich jedoch, dass die Korrelation der Fehler mit zunehmender Testlänge ansteigt. Daher dürfte die Regression zur Mitte bei kürzeren Skalen stärker ausgeprägt sein. Im Vergleich zu den anderen Verfahren dürften die multivariaten Selektionsmethoden also am besten abschneiden, wenn mehr als die Hälfte der Items in den Test aufgenommen wird.

Verteilung der Varianzkomponenten

Die Varianz der Items setzt sich aus drei verschiedenen Komponenten (valide Varianz, gemeinsame systematische Fehlervarianz, spezifische Fehlervarianz) zusammen. Die Kommunalität ist die Summe aus valider Varianz und gemeinsamer systematischer Fehlervarianz. Der Resultate der verschiedenen Selektionsstrategien dürften nicht zuletzt von der Verteilung der Itemvarianzen auf diese Komponenten abhängen:

- Valide Varianz:

  Damit ist diejenige Varianz gemeint, welche die Items mit dem Kriterium teilen. Da es in der Regel möglich ist, die Items so zu polen, dass sie positive Validitätskoeffizienten haben, dürfte der Gesamtsummenscore aller Items des Itempools umso valider sein, je höher der Anteil der validen Varianz ist. Da die Validität des Summenscores ein Moderator des Zusammenhangs zwischen Trennschärfe und Itemvalidität ist (vgl. Kapitel 2.1, S. 26), ist bei hohen Itemvaliditäten kein großer Unterschied zwischen der Selektion anhand der Trennschärfe und der Selektion anhand der Itemvalidität zu erwarten. Je höher die Itemvaliditäten sind, desto weniger dürfte die gemeinsame systematische Fehlervarianz der Items ins Gewicht fallen. Wenn die valide Varianz dominiert, müsste daher auch die Selektion der Items anhand der internen Konsistenz zu validen Skalen führen. Multivariate Verfahren der Itemselektion anhand von Validitätsdaten dürften unter diesen Bedingungen nicht wesentlich besser sein als herkömmliche Methoden der Itemselektion.

- Gemeinsame systematische Fehlervarianz

  Damit sind Varianzanteile gemeint, die zwar in mehreren Items vorkommen, die aber keinen direkten Beitrag für die Vorhersage des Kriteriums leisten. Nur bei unkorrelierten Messfehlern setzt sich die gemeinsame systematische Varianz der Items ausschließlich aus wahrer Varianz im Sinne der klassischen Testtheorie zusammen. Multivariate Verfahren der Itemselektion ermöglichen es, Skalen zu konstruieren bei denen die gemeinsame systematische Fehlervarianz möglichst gering ist (buffered scales sensu Cattell und Tsujioka, 1964). Wenn dieser Varianzanteil aber ohnehin gering ist, ist von diesen Verfahren jedoch keine substantielle Steigerung der Validität zu erwarten. Der Nutzen von multivariaten Verfahren der Itemselektion hängt jedoch nicht nur davon ab, wie groß die Ladungen der Items auf den gemeinsamen Fehlerfaktoren sind, sondern auch davon, wie die Ladungen verteilt sind. Wenn positive Ladungen genauso häufig vorkommen wie negative Ladungen, dann dürfte die gemeinsame Fehlervarianz mit zunehmender Testlänge wesentlich geringer kumulieren als die valide Varianz, sofern die Items so gepolt sind, dass sie überwiegend positive Validitätskoeffizienten haben. Bei relativ homogenen Ladungen auf den gemeinsamen Fehlerfaktoren kumuliert die gemeinsame Fehlervarianz dagegen in einem vergleichbaren Ausmaß wie die valide Varianz. Bei längeren Tests sollten multivariate Verfahren zur Maximierung der Validität bei relativ homogenen Ladungen auf den gemeinsamen Fehlerfaktoren von Vorteil sein, da sie eine Minimierung (Suppression nur möglich, wenn Vorzeichen der Ladungen variieren) der gemeinsamen Fehlervarianz ermöglichen, solange die Ladungen der Items auf den Fehlerfaktoren nicht vollkommen identisch sind.

  Auch die Anzahl der gemeinsamen Fehlerfaktoren könnte von Bedeutung sein. Wenn sich die gemeinsame, spezifische Varianz auf mehrere Faktoren verteilt, könnte es selbst bei Anwendung multivariater Methoden zur Validitätsmaximierung schwer fallen, die systematische Fehlervarianz deutlich zu reduzieren, da die Ladungen auf mehreren Faktoren gleichermaßen zu berücksichtigen sind. In der Praxis dürfte eine große Anzahl gemeinsamer Fehlerfaktoren jedoch auch mit einer höheren gemeinsamen Fehlervarianzanteil einhergehen, was die Ausnutzung von Suppressionseffekten wiederum vielversprechender erscheinen lässt. Multivariate Verfahren der Validitätsmaximierung dürften demnach besonders lohnend sein, wenn sich viel gemeinsame Fehlervarianz auf wenige gemeinsame Fehlerfaktoren konzentriert.

- Spezifische Varianz

  Damit sind solche Varianzanteile gemeint, die ein Item weder mit dem Kriterium noch mit den anderen Items teilt. Neben unsystematischer Fehlervarianz im Sinne der klassischen Testtheorie (vgl. [1.1-2] auf S. 3) können dies auch systematische Varianzanteile sein, die nur durch dieses Item abgebildet werden. In der Regel ist es nicht notwendig zwischen unsystematischen und systematischen spezifischen Varianzanteilen zu unterscheiden, da sie sich gleichermaßen mindernd auf die Korrelationen mit anderen Variablen auswirken. Der Anteil spezifischer Varianz an der Gesamtvarianz der aufsummierten Items verringert die Validität. Spezifische Varianzanteile verringern die Validität eines Tests umso mehr, je geringer die gemeinsamen Varianzanteile sind (vgl. [1.2-6] auf S. 18). Das Ausnutzen von Suppressionseffekten führt zwangsläufig dazu, dass die gemeinsame Varianz geringer ist, da bei Suppression ja irrelevante gemeinsame Varianzanteile unterdrückt werden. An sich wirkt sich dies positiv auf die Validität aus. Allerdings führt dies auch zu einem umso stärkeren Absinken der Validität durch spezifische Varianzanteile. Je mehr bei der Itemselektion also auf die Kumulation valider Varianz zugunsten der Suppression invalider gemeinsamer Varianzanteile verzichtet wird, desto stärker sollte die Minderung der Validität durch spezifische Varianzanteile zu Buche schlagen. Zufällige, unkorrelierte Messfehler könnten daher den Nutzen der Berücksichtigung von Suppressionseffekten nivellieren.

  Die Minderung der Validität durch spezifische Varianzanteile stellt auch deswegen ein Problem für multivariate Methoden der Itemselektion dar, da sie die Validitätsunterschiede der zur Auswahl stehenden Itemkombinationen verringern dürften. Dies lässt sich anhand der Verdünnungsformeln (vgl. [1.2-6] auf S. 18) nachvollziehen. Eine geringe Varianz der wahren Validitätskoeffizienten der zur Auswahl stehenden Tests führt aber dazu, dass die Regression zur Mitte durch Schätzfehler zunimmt (vgl. [2.2-3] auf S. 37 und [2.2-4] auf S. 37). Da die multivariaten Verfahren der Testkonstruktion wesentlich anfälliger für die Regression zur Mitte sein dürften, dürften sie insbesondere bei der Auswahl eines geringen Anteils der Items auf der Grundlagen von Validitätsdaten, die an kleinen Personenstichproben gewonnen wurden, keine Vorteile bringen (vgl. [2.2-5] auf S. 38).

Kommunalität des Kriteriums

Schließlich muss man berücksichtigen, dass in der Praxis auch das Kriterium mit Messfehlern behaftet ist. Die Messfehler des Kriteriums reduzieren genau wie die Messfehler der Items die wahren Validitätsunterschiede der zur Auswahl stehenden Skalen. Im Gegensatz zur spezifischen (Fehler-)Varianz der Items verringert die spezifische Varianz des Kriteriums nur die

Unterschiede zwischen den zur Disposition stehenden Itemkombination und kann demnach nicht zu Änderungen in der Rangfolge der Validität der Skalen führen. Eine geringe Kommunalität des Kriteriums mindert also lediglich die Validitätsunterschiede zwischen den zur Disposition stehenden Itemkombinationen. Der Fehler bei der Schätzung der einzelnen Validitätskoeffizienten hängt dagegen lediglich von Stichprobenumfang ab. Daher dürfte die Regression zur Mitte stärker ausgeprägt sein, je größer der Messfehler des Kriteriums ist. Allerdings betrifft dies nur solche Verfahren, die sich auf Validitätsdaten stützen. Bei multivariaten Verfahren, die eine Vielzahl von möglichen Itemkombinationen berücksichtigen, dürfte dies besonders starke Auswirkungen haben. Wenn dagegen der Stichprobenfehler bei der Ermittlung der Validitätskoeffizienten aufgrund eines großen Stichprobenumfangs gering ist, sollte eine niedrige Kommunalität des Kriteriums kaum Konsequenzen haben.

5 Simulationsstudie

Aus den obigen Ausführungen ergibt sich, dass sich die Validität aller Skalen beliebiger Länge und Zusammensetzung, die sich aus einem Itempool bilden lassen, anhand der Kovarianzmatrix der Items und des Validitätskriteriums bestimmen lässt (vgl. [2.2-2] auf S. 31). Die Effektivität verschiedener Itemselektionsalgorithmen lässt sich also allein auf Grundlage der Kovarianzmatrix beurteilen. Für verschiedene Kovarianzmatrizen könnte man jedoch jeweils zu anderen Ergebnissen kommen. Um die Frage nach dem optimalen Itemselektionsalgorithmus mit einem gewissen Grad an Allgemeingültigkeit beurteilen zu können, sollte man demnach aus der Menge aller Kovarianzmatrizen eine möglichst repräsentative Stichprobe ziehen und die Güte der Ergebnisse der einzelnen Selektionsprozeduren vergleichen.

Eine Stichprobe der Kovarianzmatrizen kann man entweder per Zufall generieren oder man greift auf Kovarianzmatrizen zurück, die anhand empirischer Testdaten geschätzt wurden. Jede der beiden Möglichkeiten hat ihre eigenen Vor- und Nachteile. Die Ergebnisse von Simulationsstudien mögen einen größeren Grad an Allgemeinheit haben, da hierbei auch Konstellationen berücksichtigt werden können, die in empirischen Anwendungen selten vorkommen. Des Weiteren kann in Simulationsstudien eine hinreichend große Zahl von Kovarianzmatrizen erzeugt werden, so dass die Ergebnisse nicht von den spezifischen Eigenschaften eines bestimmten Itempools und einer bestimmten Stichprobe von Personen abhängen und daher weniger zufallsbehaftet sind. Außerdem können in einer Simulationsstudie auch die Anzahl und Art von Varianz- und Kovarianzquellen explizit festgelegt werden und theoretisch interpretiert werden. Schließlich können die simulierten Kovarianzmatrizen als fehlerfreie Populationswerte interpretiert werden.

Andererseits ist die externe Validität von Simulationsstudien fraglich, da die Kovarianzmatrizen empirischer Testdaten eben keine Zufallsstichprobe aus der Menge aller Kovarianzmatrizen sind, sondern im Allgemeinen die Zusammenhänge zwischen Variablen (bzw. Items) darstellen, die alle mit der Intention entwickelt wurden, dasselbe Merkmal zu messen, so dass sie sich durch spezifische Eigenschaften (insbesondere ein gewisses Maß an Homogenität) auszeichnen dürften, die u.U. in den Simulationsstudien nicht angemessen modelliert wurden. Im Rahmen dieser Arbeit werden daher neben den Simulationsstudien auch Vergleiche der Selektionsmethoden anhand empirischer Testdaten untersucht. In diesem Kapitel werden die Selektionsstrategien anhand zufällig generierter Kovarianzmatrizen verglichen, während im nächsten Kapitel auf empirische Testdaten zurückgegriffen wird.

5.1 Methode

In der Praxis ist die Kovarianzmatrix des Itempools nicht bekannt, sondern sie muss anhand empirischer Daten geschätzt werden. Es ist plausibel, dass die verschiedenen Selektionsverfahren unterschiedlich robust auf Fehler bei der Schätzung der wahren Kovarianzen reagieren. Daher reicht es nicht aus, Kovarianzmatrizen zu simulieren, welche die wahren Zusammenhänge der Variablen beschreiben, sondern es ist notwendig, Schätzfehler in die Simulationsstudie einzubeziehen. Da die Konsequenzen der Schätzfehler für beliebige Kovarianzmatrizen kaum analytisch beurteilt werden können (vgl. Kapitel 2.2.4), bietet es sich an, auch hier auf Simulationen zurückzugreifen. Dazu muss ein Zufallsalgorithmus gefunden werden, der aus einer Personenpopulation mit bekannter Kovarianzmatrix zufällige (Personen-)Stichproben des gewünschten Umfangs zieht. Die Kovarianzmatrix der Stichprobe wird dabei umso weniger von derjenigen der Population abweichen, je größer die Anzahl der (simulierten) Testpersonen ist. Die Simulation der Stichprobenkovarianzmatrix dient also ausschließlich dazu, den Einfluss von Fehlern bei der Schätzung der wahren Kovarianz auf die Effektivität der einzelnen Verfahren zu berücksichtigen. Die weiter oben skizzierte Simulation der Kovarianzmatrix in der Population soll dagegen den Nutzen der Selektionsverfahren in Abhängigkeit von den statistischen Eigenschaften des Itempools analysieren. Der Ablauf der Simulationsstudien ist in Tabelle 2 skizziert. Im Folgenden sollen die in der Simulationsstudie angewandeten Methoden näher geschildert werden.

Tabelle 2: Ablauf der Simulationsstudien

1. Simulation der Kovarianzmatrix der Population.
2. Simulation der Kovarianzmatrix in der Stichprobe
3. Itemselektion anhand der Stichproben-Kovarianzmatrix.
4. Maße der Testgüte der selektierten Subtests in der Population berechnen.

→ Zurück zu 1.

5.1.1 Generierung von Kovarianzmatrizen

Bei der Simulation von Kovarianzmatrizen sind zwei Dinge zu beachten. Zum einen sollte jede beliebige Kovarianzmatrix mit einer von null verschiedenen Wahrscheinlichkeit(-sdichte) in die Studie miteinbezogen werden. Andererseits sollte ausgeschlossen werden, dass Matrizen in die Studie Eingang finden, die gar keine Kovarianzmatrizen sein können. Daher müssen zunächst

hinreichende und notwendige Bedingungen dafür gefunden werden, dass eine Matrix die Kovarianzmatrix reeller Zufallsvariablen sein könnte.

Bereits an der Definition der Kovarianz lässt sich unmittelbar die Symmetrie als erste notwendige Bedingung einer Kovarianzmatrix erschließen. Gegeben sei ein Zufallsvektor **x**, in dessen Zeilen *n* beliebige reellen Zufallsvariablen ($x_1, \ldots, x_n$) eingetragen sind. Für die Kovarianz jedes Variablenpaares gilt (vgl. Bortz, 1999):

$$\sigma\left(x_i, x_j\right) = \sigma\left(x_j, x_i\right) \qquad i,j \in \{1, \ldots, n\}$$

[5.1-1]

Hieraus ergibt sich direkt die Symmetrie der Kovarianzmatrix Σ_x:

$$\Sigma_{\mathbf{x}} := \begin{pmatrix} \sigma(x_1,x_1) & \sigma(x_1,x_2) & \cdots & \sigma(x_1,x_n) \\ \sigma(x_2,x_1) & \sigma(x_2,x_2) & \cdots & \sigma(x_2,x_n) \\ \vdots & & & \\ \sigma(x_n,x_1) & \sigma(x_n,x_j) & \cdots & \sigma(x_n,x_n) \end{pmatrix},$$

[5.1-2]

wobei in der Diagonalen jeweils die Varianzen der Variablen stehen. Aus der Kovarianzmatrix lassen sich jedoch nicht nur die Kovarianzen der ursprünglichen Zufallsvariablen, auf die sie sich bezieht, ermitteln, sondern auch die Kovarianzen beliebiger Linearkombinationen dieser Zufallsvariablen.

$$\sigma(\mathbf{a'x}, \mathbf{b'x}) = \sigma\left(\sum_{i=1}^{n} a_i x_i, \sum_{j=1}^{n} b_j x_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \sigma(x_i, x_j) = \mathbf{a'}\Sigma_{\mathbf{x}}\mathbf{b} \qquad (\mathbf{a},\mathbf{b} \in \mathbb{R}^n)$$

[5.1-3]

$a_1, \ldots, a_n, b_1, \ldots, b_n$ sind beliebige reelle Zahlen, während $x_1, \ldots, x_n$ beliebige Zufallsvariablen sind.

Da die Kovarianz jeder reellen Zufallsvariable mit sich selbst (Varianz) größer oder gleich null ist, ist Σ_x positiv semi-definit, falls es eine Linearkombination **a'x** der ursprünglichen Zufallsvariable gibt, deren Varianz $\mathbf{a'}\Sigma_x\mathbf{a}$ gleich null ist und positiv definit, falls es kein $\mathbf{a} \in R^n\backslash\{0\}$ gibt, so dass $\mathbf{a'}\Sigma_x\mathbf{a} = 0$ (Graybill, 1961, S. 3). Die positive Definitheit oder Semi-Definitheit ist die zweite notwendige Bedingung, die eine Kovarianzmatrix erfüllen muss.

Um zu zeigen, dass die positive Definitheit oder Semi-Definitheit einer symmetrischen Matrix auch eine hinreichende Bedingung dafür ist, dass sie eine Kovarianzmatrix sein kann, wird im Folgenden demonstriert, wie sich zu jeder positiv definiten symmetrischen Matrix **K**

Zufallsvariablen definieren lassen, deren Kovarianzmatrix gleich **K** ist. Dazu wird zunächst erläutert, in welcher Beziehung die Kovarianzmatrix einer Menge von Zufallsvariablen zu den Kovarianzmatrizen einer Menge von Linearkombinationen dieser Zufallsvariablen stehen.

Gegeben seien *n* beliebige Zufallsvariablen. Die Menge aller Linearkombinationen der ursprünglichen Zufallsvariablen spannt einen Vektorraum V von reellen Zufallsvariablen auf. Durch die Kovarianz ist ein Skalarprodukt[8] auf V gegeben, das jedem Paar von Zufallsvariablen ihre Kovarianz zuordnet (S: V×V→$\mathbb{R}$; $v,w \mapsto \sigma(v,w)$ für alle $v,w \in$ V, vgl. Fischer, 1993).

Gegeben sei ein p-dimensionaler Zufallsvektor $\mathbf{v} = \mathbf{A}'\mathbf{x} = \begin{pmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{a}_1'\mathbf{x} \\ \vdots \\ \mathbf{a}_p'\mathbf{x} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} a_{1i}x_i \\ \vdots \\ \sum_{i=1}^{n} a_{pi}x_i \end{pmatrix}$, in dessen

Zeilen Zufallsvariablen aus V eingetragen sind ($\mathbf{a}_k \in \mathbb{R}^n$, $\mathbf{a}_k'\mathbf{x} \in$ V). Nach [5.1-3] ergibt sich für die Kovarianz der *k*-ten mit der *l*-ten Zeile von **v**:

$$\sigma(v_k,v_l) = \mathbf{a}_k' \, \Sigma_{\mathbf{x}} \mathbf{a}_l \qquad \textbf{[5.1-4]}$$

Die Kovarianzmatrix von **v** ist demnach:

$$\Sigma_{\mathbf{v}} = \begin{pmatrix} \mathbf{a}_1'\Sigma_{\mathbf{x}}\mathbf{a}_1 & \mathbf{a}_1'\Sigma_{\mathbf{x}}\mathbf{a}_2 & \cdots & \mathbf{a}_1'\Sigma_{\mathbf{x}}\mathbf{a}_p \\ \mathbf{a}_2'\Sigma_{\mathbf{x}}\mathbf{a}_1 & \mathbf{a}_2'\Sigma_{\mathbf{x}}\mathbf{a}_1 & \cdots & \mathbf{a}_2'\Sigma_{\mathbf{x}}\mathbf{a}_p \\ \vdots & & & \\ \mathbf{a}_p'\Sigma_{\mathbf{x}}\mathbf{a}_1 & \mathbf{a}_p'\Sigma_{\mathbf{x}}\mathbf{a}_1 & \cdots & \mathbf{a}_p'\Sigma_{\mathbf{x}}\mathbf{a}_p \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix} \begin{pmatrix} \Sigma_{\mathbf{x}}\mathbf{a}_1 & \cdots & \Sigma_{\mathbf{x}}\mathbf{a}_p \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix} \Sigma_{\mathbf{x}} \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_p \end{pmatrix} = \mathbf{A}'\Sigma_{\mathbf{x}}\mathbf{A}$$

[5.1-5]

Gegeben sei nun eine positiv definite oder positiv semi-definite symmetrische Matrix **K**. Wegen der Symmetrie von **K,** gibt es Eigenvektoren ($\mathbf{c}_1, \ldots, \mathbf{c}_n \in \mathbb{R}^n$) von **K**, die eine Orthonormalbasis des $\mathbb{R}^n$ bilden (Fischer, 1993, S.206). Aufgrund der Orthogonalität der Matrix $\mathbf{C} := \begin{pmatrix} \mathbf{c}_1 & \cdots & \mathbf{c}_1 \end{pmatrix}$ gilt (Fischer, 1993, S.201):

[8] Da in der linearen Algebra Normen und Winkel über Skalarprodukte definiert werden, lassen sich so auch Normen und Winkel zwischen den Zufallsvariablen aus V definieren. Die Varianz ist eine Norm auf V und der "Winkel" zwischen zwei Zufallsvariablen aus V entspricht dem Arcuscosinus der Korrelation, wie sich leicht nachprüfen lässt.

$\mathbf{C}^{-1} = \mathbf{C}'$.

[5.1-6]

Daher ergibt sich:

$$\mathbf{C}^{-1}\mathbf{K}\mathbf{C} = \mathbf{C}'\mathbf{K}\mathbf{C} = \begin{pmatrix} \mathbf{c}_1' \\ \vdots \\ \mathbf{c}_n' \end{pmatrix} \mathbf{K} \begin{pmatrix} \mathbf{c}_1 & \cdots & \mathbf{c}_n \end{pmatrix} = \begin{pmatrix} \mathbf{c}_1' \\ \vdots \\ \mathbf{c}_n' \end{pmatrix} \begin{pmatrix} \lambda_1 \mathbf{c}_1 & \cdots & \lambda_n \mathbf{c}_n \end{pmatrix} = \begin{pmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n \end{pmatrix} =: \mathbf{\Lambda}, \quad (5)$$

[5.1-7]

wobei λ_i der Eigenwert ist, der dem Eigenvektor $\mathbf{c}_i$ zugeordnet ist. Da **K** positiv (semi-)definit ist, sind alle Eigenwerte größer (oder gleich) null (Fischer, 1993, S. 214). Es sei nun ein Zufallsvektor $\mathbf{x'} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}$ gegeben, in dessen Spalten *n* unkorrelierte Zufallsvariablen stehen, deren Varianz den Eigenwerten von K entspricht ($\sigma^2(x_i)=\lambda_i$). Die Diagonalmatrix $\mathbf{\Lambda}$ ist offensichtlich die Kovarianzmatrix $\mathbf{\Sigma_x}$ der Zufallsvariablen x_1 , . . . , x_n . Unter Rückgriff auf [5.1-5], [5.1-6] und [5.1-7] errechnet sich die Kovarianzmatrix des Zufallsvektors $\mathbf{y} = \mathbf{C}\mathbf{x}$ wie folgt:

$$\Sigma_{\mathbf{y}} = \Sigma_{\mathbf{Cx}} = \mathbf{C}\Sigma_{\mathbf{x}}\mathbf{C}' = \mathbf{C}\Lambda\mathbf{C}^{-1} = \mathbf{C}\mathbf{C}^{-1}\mathbf{K}\mathbf{C}\mathbf{C}^{-1} = \mathbf{K}$$

[5.1-8]

Die Matrix **K** kann also offensichtlich eine Kovarianzmatrix sein. Damit wäre gezeigt, dass Symmetrie und positive Definitheit bzw. Semi-Definitheit notwendige und hinreichende Bedingungen, dafür sind, dass eine Matrix Kovarianzmatrix sein könnte.

Um Kovarianzmatrizen einer gegeben Dimension *n* zu simulieren, muss jetzt also nur noch ein Zufallsverfahren gefunden werden, das einerseits ausschließlich positiv (semi-)definite Matrizen generiert und anderseits in der Lage ist, jede *n*-dimensionale Kovarianzmatrix (mit positiver Wahrscheinlichkeitsdichte) zu generieren.

Das Produkt jeder Matrix mit ihrer Transponierten ist positiv (semi-)definit (Graybill, 1961, S. 4). Daher bietet es sich an, einen Zufallsalgorithmus zu wählen, bei dem jede Kombination von reellen Zahlen in den Zellen einer Matrix vorkommen kann (z.B. multivariate Normalverteilung). Multipliziert man diese Matrix mit ihrer Transponierten, so erhält man eine symmetrische positiv (semi-)definite Matrix, was gleichbedeutend mit der Aussage ist, dass diese Matrix eine Kovarianzmatrix sein könnte. Will man sicherstellen, dass jede nur denkbare Kovarianzmatrix als Ergebnis dieses Zufallsalgorithmus resultieren könnte, so ist zu zeigen, dass

sich jede positiv (semi-)definite symmetrische Matrix **K** als Produkt einer anderen Matrix mit ihrer Transponierten darstellen lässt. Nach [5.1-7] gilt: $\mathbf{C}'\mathbf{K}\mathbf{C} = \mathbf{\Lambda}$. Umformen ergibt[9]:

$$\mathbf{K} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}' = \mathbf{C}\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}\mathbf{C}' = \begin{pmatrix}\mathbf{c}_1 & \cdots & \mathbf{c}_n\end{pmatrix}\begin{pmatrix}\lambda_1^{1/2} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n^{1/2}\end{pmatrix}\begin{pmatrix}\lambda_1^{1/2} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n^{1/2}\end{pmatrix}\begin{pmatrix}\mathbf{c}_1' \\ \vdots \\ \mathbf{c}_n'\end{pmatrix}$$

$$= \begin{pmatrix}\lambda_1^{1/2}\mathbf{c}_1 & \cdots & \lambda_n^{1/2}\mathbf{c}_n\end{pmatrix}\begin{pmatrix}\lambda_1^{1/2}\mathbf{c}_1' \\ \vdots \\ \lambda_n^{1/2}\mathbf{c}_n'\end{pmatrix} = \mathbf{C}\mathbf{\Lambda}^{1/2}(\mathbf{C}\mathbf{\Lambda}^{1/2})' =: \mathbf{L}\mathbf{L}'$$

[5.1-9]

Für jede Kovarianzmatrix gibt es also (mindestens) eine Möglichkeit sie als das Produkt einer anderen Matrix mit ihrer Transponierten darzustellen. Trägt man in die Zellen einer $n \times n$-dimensionalen Matrix **L** unabhängige standardnormalverteilte Zufallsvariablen ein, so kann jede beliebige Kombination reeller Zufallsvariablen (mit positiver Wahrscheinlichkeitsdichte) vorkommen. Nach [5.1-9] folgt daraus, dass auch jede n-dimensionale Kovarianzmatrix (mit positiver Wahrscheinlichkeitsdichte) vorkommen kann. Der hier skizzierte Zufallsalgorithmus ist also in der Lage jede beliebige Kovarianzmatrix zu generieren. Allerdings werden nicht alle Kovarianzmatrizen mit derselben Wahrscheinlichkeit(-sdichte) generiert. Die Kovarianzmatritzen, die durch diesen Algorithmus generiert werden, folgen einer Wishart-Verteilung mit n Freiheitsgraden und der Einheitsmatrix als Skalierungsfaktor. Die höchste Wahrscheinlichkeitsdichte ergibt sich demnach für die Einheitsmatrix.

Stichproben von Kovarianzmatrizen, die eher repräsentativ für empirische Datensätze sind, lassen sich unter Rückgriff auf das Faktormodell der Persönlichkeit entwickeln. Auch der bereits dargestellte Zufallsalgorithmus lässt sich im Sinne des Faktormodells interpretieren. Die Einträge f_{ij} in den Spalten der Matrix **L** lassen sich nämlich als das Ausmaß der Beeinflussung (Faktorladung) durch unabhängige normierte ($\sigma^2 = 1$) Varianzquellen (Faktoren) $z_1, \ldots, z_n$ auffassen:

$$\mathbf{y}=\mathbf{L}\mathbf{z} := \mathbf{L}(z_1 \ldots z_n)'$$

[5.1-10]

Nach [5.1-5] (S. 59) gilt dann:

[9] Im Gegensatz zur bekannten Cholesky Dekomposition (vgl. SAS/IML Usage and Reference, 1989, S. 411; Graybill, 1961, S.3) bezieht sich [5.1-9] auch auf singuläre Kovarianzmatrizen.

$$\mathbf{\Sigma_y = L\Sigma_z L' = L1L' = LL' = K}\ .$$

[5.1-11]

Falls keiner der Eigenwerte null ist, lassen sich die Faktoren auch als Linearkombination der ursprünglichen Variablen darstellen, indem man [5.1-10] von links mit $\mathbf{\Lambda^{-1/2}C'}$ multipliziert ($\mathbf{\Lambda^{-1/2}}$ ist die zu $\mathbf{\Lambda^{1/2}}$ inverse Matrix):

$$\mathbf{\Lambda^{-1/2}C'y = \Lambda^{-1/2}C'Lz = \Lambda^{-1/2}C'C\Lambda^{1/2}z = \Lambda^{-1/2}C^{-1}C\Lambda^{1/2}z = \Lambda^{-1/2}\Lambda^{1/2}z = z}$$

[5.1-12]

Gibt es Eigenwerte $\lambda_i = 0$, so stehen in der entsprechenden Spalte der Matrizen **L** und $\mathbf{\Lambda^{1/2}}$ nur Nullen. Entfernt man die entsprechenden Spalten aus den Matrizen **L**, **C** und $\mathbf{\Lambda^{1/2}}$ sowie die entsprechenden Faktoren aus dem Zufallsvektor **z**, so ändert sich an dem Produkt **LL'** nichts. In diesem Fall existiert $\mathbf{\Lambda^{-1/2}}$ und die Faktoren lassen sich wieder über [5.1-12] als Linearkombination der ursprünglichen Variablen darstellen.

Im Faktormodell der Persönlichkeit werden jedoch im Allgemeinen mehr latente Varianzquellen zugrundegelegt als manifeste Variablen existieren, da neben den gemeinsamen Faktoren für jede einzelne Variable je ein zusätzlicher spezifischer Faktor (sowie ein Fehlerfaktor) postuliert wird (vgl. [1.1-19] auf S. 11). In diesem Fall ist es nicht mehr möglich, die Faktoren als Linearkombination der manifesten Variablen darzustellen. Die manifesten Variablen sind jedoch gemäß [5.1-10] nach wie vor Linearkombinationen der latenten Variablen (gemeinsame, spezifische und Fehlerfaktoren). In empirischen Anwendung macht es meist wenig Sinn, zwischen spezifischer und Fehlervarianz zu unterscheiden, da sie sich in der Regel nicht separat schätzen lassen. In der vorliegenden Simulationsstudie macht dies auch konzeptuell keinen Sinn, da die externe Validität der spezifischen Faktoren der einzelnen Items per definitionem gleich null sein muss, wenn auch das Kriterium in der Menge der manifesten Variablen enthalten ist. Gemäss des Faktormodells ergibt sich demnach folgende allgemeine Struktur der Ladungsmatrix **L**:

$$\mathbf{L} = \begin{pmatrix} l_{11} & \cdots & l_{1n} & \varepsilon_1 & & \mathbf{0} \\ \vdots & \ddots & \vdots & & \ddots & \\ l_{n1} & \cdots & l_{nn} & \mathbf{0} & & \varepsilon_n \end{pmatrix} \qquad l_{ij} = 0 \text{ für } j > N_F$$

[5.1-13]

N_F ist die Anzahl der gemeinsamen Faktoren; l_{ij} ist die Ladung der manifesten Variable i auf dem Faktor j; ε_i ist die Ladung der Variablen i auf dem zugehörigen Fehlerfaktor. Mit Hilfe der

Matrix **L** lassen sich alle bisher besprochenen Modelle über den Zusammenhang von manifesten und latenten Variablen modellieren**.**

Wenn alle Parameter frei variieren dürfen, so entspricht dies dem Modell der Faktorenanalyse. Für die vorliegende Simulationsstudie ist es – so wie in den meisten empirischen Anwendungen – jedoch sinnvoll, wenn die Anzahl der gemeinsamen Faktoren N_F deutlich kleiner ist als die Anzahl der manifesten Variablen *n.* Wenn man alle ε_i auf null setzt, und fordert, dass sämtliche von null verschiedenen Spaltenvektoren linear unabhängig sind, so sind die latenten Faktoren als Hauptkomponenten zu interpretieren (Basilevsky, 1994). Da sich für jede Kovarianzmatrix Hauptkomponenten berechnen lassen, kann auf diesem Wege jede nur erdenkliche Datenstruktur modelliert werden. Setzt man dagegen N_F auf 1 und lässt nur positive Ladungen auf dem einzigen gemeinsamen Faktor zu, so entspricht dies dem Modell τ-kongenerischer Variablen (vgl. [1.1-17] auf S. 10). Wenn man zudem fordert, dass alle manifesten Variablen identische Ladungen auf dem ersten Faktor haben, so entspricht dies dem Modell essentiell τ-äquivalenter Variablen (vgl. [1.1-16] auf S. 10). Wenn auch die Ladungen auf den Fehlerfaktoren identisch sind, dann entspricht dies dem Modell essentiell τ-äquivalenter Variablen mit homogener Fehlervarianz. Wenn man nun für diejenigen Parameter, die in den verschiedenen Modellen noch frei variieren dürfen, jeweils standardnormalverteilte Zufallsvariablen einsetzt, dann können sich alle Datenstrukturen ergeben, die bei Geltung der jeweiligen Modelle möglich sind, da sich für jede denkbare Faktorladungsmatrix eine positive Wahrscheinlichkeitsdichte ergibt.

[5.1-13] bietet demnach die Möglichkeit, die Simulationsstudie so anzulegen, dass die per Zufall generierten Parameter theoretisch bedeutsam und inhaltlich interpretierbar sind, wobei der gewählte Algorithmus gleichzeitig so flexibel ist, dass jede nur denkbare Konstellation berücksichtigt wird. Es muss betont werden, dass bisher weder für die manifesten noch für die latenten personenbezogenen Variablen Verteilungsannahmen gemacht wurden. Die Standardnormalverteilung wurde lediglich gewählt, um einen Zufallsalgorithmus zu konstruieren, der jede Kovarianzmatrix, die sich bei Geltung der jeweiligen Modellannahmen ergeben könnte, hervorbringen kann. Wenn die Parameter dieses Zufallsalgorithmus multivariat normalverteilt sind, kommt allerdings nicht jede Kovarianzmatrix mit gleicher Wahrscheinlichkeit(sdichte) vor.

Ferner ist zu beachten, dass die manifesten Zufallsvariablen **y** hier nicht z-standardisiert wurden, wie bei der Faktorenanalyse üblich. Daher entsprechen die Einträge in der Matrix **L** nicht wie gewohnt Korrelationen, sondern der Kovarianz des Faktors mit der entsprechenden

Zufallsvariable. Die Kovarianz von manifester Variable und latentem Faktor, ist hier das Produkt aus deren Korrelation und der Standardabweichung der manifesten Variable, da die Varianz der Faktoren jeweils auf 1 normiert ist. Die Summe der quadrierten Ladungen der manifesten Variablen auf den latenten Faktoren ist nicht 1 (oder eine Kommunalitätsschätzung), sondern die Varianz der Variablen. Da die Faktoren normiert und unkorreliert sind, ist ihre Kovarianzmatrix die Einheitsmatrix.

5.1.2 Ziehung der Personenstichprobe

Im vorangegangenen Abschnitt wurde ein Algorithmus beschrieben, mit dem sich jede beliebige Kovarianzmatrix unter Rückgriff auf das Faktorenmodell der Persönlichkeit simulieren lässt. Die simulierten Kovarianzmatrizen beschreiben denkbare statistische Zusammenhänge zwischen den Items eines Itempools sowie mit dem Kriterium. Die Kovarianzmatrix gibt die statistischen Eigenschaften der Items in einer (hypothetischen) Personenpopulation wieder. Anhand der Kovarianzmatrix der Items und des Kriteriums lässt sich die Validität jedes Subtests leicht ermitteln (vgl. [2.2-2] auf S. 31). In der Praxis ist diese Kovarianzmatrix jedoch nicht bekannt, sondern sie muss, anhand empirischer Daten geschätzt werden. Die damit verbundenen Fehler könnten die Ergebnisse der verschiedenen Verfahren zur Itemselektion in unterschiedlicher Weise beeinträchtigen. Um in der vorliegenden Simulationsstudie auch die Einflüsse von Schätzfehlern zu berücksichtigen, ist es notwendig, auch die Ziehung einer Personenstichprobe zu simulieren.

Der im vorigen Abschnitt beschriebene Algorithmus zur Generierung von Kovarianzmatrizen beruht darauf, dass die Matrix **L** mit ihrer Transponierten multipliziert wird. Wie bereits erwähnt, lassen sich die Einträge in der Matrix **L** als Ausmaß der Beeinflussung durch unkorrelierte standardisierte Faktoren auffassen (vgl. [5.1-10] auf S. 61). Bei zufälliger Ziehung (der Faktorwerte) aus beliebigen, unkorrelierten, standardisierten Verteilungen ($z_1, \ldots, z_n$) ergibt sich folglich nach Multiplikation des Vektors $\mathbf{z}=(z_1 \ldots z_n)'$mit der Matrix **L** ein Zufallsvektor **y**, dessen Kovarianzmatrix $\mathbf{\Sigma}_y$ der simulierten Kovarianzmatrix **L'L** entspricht. Man muss demnach lediglich Faktorwerte der (hypothetischen) Personen per Zufallsgenerator simulieren und anschließend durch Multiplikation mit der Matrix **L** transformieren, um Stichprobendaten mit den gewünschten Eigenschaften zu simulieren. Wenngleich man für die Faktorwerte beliebige Verteilungen unterstellen könnte, bietet es sich an, (unabhängige) Standardnormalverteilungen zu wählen. In diesem Fall sind auch die Werte der Personen auf den Items und dem Kriterium multivariat normalverteilt. Die Stichprobenkovarianzmatrizen folgen dann einer Wishart-Verteilung mit der Skalierungsmatrix **L'L**.

Bei dem bisher geschilderten Vorgehen wäre es durchaus denkbar, dass ein großer Teil der Items negative Validität hat. Bei der Testkonstruktion wird man die Items jedoch in der Regel so polen, dass sie eine positive Validität haben[10]. Daher wurden die Items so gepolt, dass sie in der Stichprobe eine positive Validität haben. Dies bedeutet jedoch nicht unbedingt, dass auch alle Itemvaliditäten in der Population positiv sind.

Im Gegensatz zur Simulierung der Populationskovarianzmatrizen müssen bei der Simulierung der Stichprobenkovarianzmatrizen Annahmen über die Verteilung der Items und des Kriteriums gemacht werden. Daher ist bei der Generalisierung der Ergebnisse der vorliegenden Simulationsstudie auf nicht-normalverteilte Daten Vorsicht angebracht.

5.1.3 Simulation von Faktorladungsmatrizen

In den beiden vorangegangenen Abschnitten wurde beschrieben, wie Kovarianzmatrizen mit Hilfe von Faktorladungsmatrizen generiert werden. Bei diesem Algorithmus sind eine Reihe von Parametern festzulegen, wie etwa die Anzahl der Faktoren, die Varianzaufklärung der einzelnen Faktoren und die Kommunalitäten der einzelnen Variablen. In diesem Abschnitt soll nun dargelegt werden, wie diese Parameter im Rahmen der vorliegenden Studie gesetzt wurden. Ziel ist es, die Parameter so zu wählen, dass die resultierenden Kovarianzmatrizen möglichst repräsentativ für die Testkonstruktion sind.

Insgesamt wurden 3 Simulationsstudien durchgeführt, wobei sich die dritte Simulationsstudie ihrerseits in drei unterschiedliche Teile untergliedern lässt.

1. In der ersten Studie wurden die Kovarianzmatrizen so simuliert, dass alle nur denkbaren Datenstrukturen (mit positiver Wahrscheinlichkeitsdichte) vorkommen können (Studie 1). Dabei wurde auf die Einführung von Fehlerfaktoren verzichtet. Stattdessen wurden genauso viele Faktoren wie Variablen eingeführt. Die Faktoren entsprechen in diesem Fall den Hauptkomponenten der unstandardisierten Variablen. Da die Berechnung von Hauptkomponenten an keinerlei Modellannahmen gebunden ist (Basilevsky, 1994,), lassen sich so alle nur denkbaren Datenstrukturen modellieren.

[10] Es ist zwar durchaus denkbar, dass sich ein Item aufgrund von Suppressionseffekten günstiger auf die Validität auswirkt, wenn es so gepolt wurde, dass die Itemvalidität negativ ist. Die Berücksichtigung von (allen möglichen) Umpolungen führt jedoch dazu, dass der Rechenaufwand der multivariaten Verfahren exorbitant wird.

2. In der zweiten Studie wurde untersucht, wie die Selektionsalgorithmen bei einem eindimensionalen Itempool (= τ-kongenerische Items vgl. [1.1-17] auf S. 10) abschneiden (Studie 2).

3. In der dritten Studie wurde die Anzahl der gemeinsamen Faktoren variiert (Studie 3). Damit sollten die verschiedenen Selektionsalgorithmen bei Datenkonstellationen verglichen werden, die sich durch das Faktormodell der Persönlichkeit beschreiben lassen. Dabei wurden die Anteile von valider Varianz, gemeinsamer Fehlervarianz und spezifischer Varianz variiert (Studie 3a, 3b und 3c).

Die Setzung der Parameter ist in Tabelle 3 zusammengefasst.

Tabelle 3: Bestimmung der Faktorladungen in den Simulationsstudien

Studie	1	2	3a	3b	3c
Ladungen:					
Valider Faktor	$N(0,1)$	$N(0,1)$	$N(0,1)$	$N(0,f-1)$	$N(0,f-1)$
Gemeinsame Fehlerfaktoren	$N(0,1)$	–	$N(0,1)$	$N(0,1)$	$N(0,1)$
Spezifischer Fehlerfaktor	–	1	$\sqrt{f}$	$\sqrt{2f-2}$	$\sqrt{f-1}$
Varianzkomponenten:					
Valide Varianz	$\chi^2_{df=1}$	$\chi^2_{df=1}$	$\chi^2_{df=1}$	$(f-1)\,\chi^2_{df=1}$	$(f-1)\,\chi^2_{df=1}$
Gemeinsame Fehlervarianz	$\chi^2_{df=n}$	0	$\chi^2_{df=f-1}$	$\chi^2_{df=f-1}$	$\chi^2_{df=f-1}$
Spezifische Varianz	0	1	f	$2\cdot(f-1)$	$f-1$

f: Anzahl der gemeinsamen Faktoren. *n:* Anzahl der Items im Itempool

Die Ladungen der Items auf den gemeinsamen Faktoren wurden in den Studien 1, 2 und 3a per Zufallsgenerator durch zufällige Ziehung aus einer Standardnormalverteilung ($N(0,1)$) bestimmt. Die Verteilung der validen Varianzanteile der verschiedenen Items entspricht demnach einer Zufallsauswahl aus einer χ^2-Verteilung mit einem Freiheitsgrad. Die gemeinsame Fehlervarianz der einzelnen Items ist ebenfalls eine χ^2-verteilte Zufallsvariable, wobei die Freiheitsgrade der Anzahl der gemeinsamen Fehlerfaktoren entsprechen.

Diese Setzung der Parameter führt jedoch in Studie 3a dazu, dass der Anteil der validen Varianz im Verhältnis zur gemeinsamen Fehlervarianz mit zunehmender Anzahl der gemeinsamen

Fehlerfaktoren abnimmt. Diese Konfundierung mag zwar der externen Validität der simulierten Datensätze dienen, da man davon ausgehen kann, dass auch bei empirischen Datensätzen der Anteil der validen Varianz bei vielen gemeinsamen Fehlerfaktoren geringer ausfällt. Die inhaltliche Interpretation der erzielten Ergebnisse wird jedoch erschwert, da die interne Validität der Ergebnisse fragwürdig ist.

Daher wurde in den Studien 3b und 3c versucht sicherzustellen, dass der relative Anteil der validen und der gemeinsamen Fehlervarianz der Items nicht von der Anzahl der gemeinsamen Fehlerfaktoren abhängt. Um zu vermeiden, dass die Itemvalidität mit der Anzahl der gemeinsamen Faktoren konfundiert ist, wurden die Ladungen auf dem validen Faktor durch Ziehung aus einer Normalverteilung bestimmt, deren Varianz der Anzahl der gemeinsamen Faktoren entspricht. Die valide Varianz entspricht dann unabhängig von der Anzahl der gemeinsamen Fehlerfaktoren im Durchschnitt der gemeinsamen Fehlervarianz der Items. Der valide Varianzanteil der Items ist also eine Zufallsvariable, deren Verteilung erst nach Division durch die Wurzel der Anzahl der gemeinsamen Fehlerfaktoren der χ^2-Verteilung mit einem Freiheitsgrad entspricht.

Demnach sind nur die Mittelwerte der beiden Varianzanteile identisch, während die Varianz und die Schiefe des validen Varianzanteils größer sind als bei der gemeinsamen Fehlervarianz, sofern mehr als ein gemeinsamer Fehlerfaktor vorliegt. Dass die Varianz und die Schiefe des Varianzanteils, der auf die gemeinsamen Fehlerfaktoren zurückgeht, in Relation zu der Varianz und Schiefe des validen Varianzanteils abnimmt, ist jedoch eine natürliche Konsequenz des Umstands, dass sich die gemeinsame Fehlervarianz auf mehrere Fehlerfaktoren verteilt. Daher wurde in der Simulationsstudie darauf verzichtet, auch die Varianz und Schiefe der beiden gemeinsamen Varianzteile zu kontrollieren, obwohl durchaus Konsequenzen auf die resultierenden Ergebnisse zu erwarten waren. Die validen Varianzanteile der einzelnen Items erreichen nämlich wesentlich häufiger extrem kleine und extrem große Werte, während die gemeinsame Fehlervarianz mit zunehmender Anzahl der gemeinsamen Fehlerfaktoren homogener über die Items verteilt ist. Dies hat zur Folge, dass die Kumulation valider Varianzteile durch die Addition der Items mit der größten validen Varianz eher zu deutlichen Steigerungen der Testvalidität führen dürfte als die Suppression der gemeinsamen Fehlervarianz. Mit zunehmender Anzahl der gemeinsamen Fehlerfaktoren dürfte bei Konstanthalten der relativen (durchschnittlichen) Anteile von valider Varianz sowie spezifischer und gemeinsamer Fehlervarianz kein besseres Abschneiden der multivariaten Verfahren zur Validitätsmaximierung zu erwarten sein.

In Studie 2 und 3a, 3b und 3c wurde jeweils eine homogene spezifische (Fehler-)Varianz der Items zugrundegelegt. Dazu wurde als Ladung auf dem spezifischen Faktor bei allen Items jeweils der gleiche Wert festgelegt. Auf eine Variation der Fehlervarianz zwischen den Items wurde verzichtet, um unrealistisch große Itemvaliditäten zu vermeiden. In Studie 1, 3a und 3b wurde die Fehlervarianz der Items so festgelegt, dass sie im Mittel genauso groß ist wie die Summe aus valider Varianz und gemeinsamer Fehlervarianz. Die Entscheidung für einen Fehleranteil der Variablen von etwa 50% beruhte auf dem Befund, dass die Retestreliabilität der Items des PRF in einer Stichprobe von N=95 bei etwa .5 lag (Amelang, Schäfer & Yousfi, 2002). Da der Nutzen der multivariaten Verfahren jedoch mitunter von der Kommunalität der Items abhängen dürfte, wurde in Studie 3c eine höhere Kommunalität der Items zugrundegelegt. Dazu wurde die Fehlervarianz so festgelegt, dass sie im Mittel nur halb so groß ist wie die Summe aus valider Varianz und gemeinsamer Fehlervarianz (Kommunalität ca. 66%).

Der Umfang des Itempools wurde in jeder der Studien jeweils in vier Schritten variiert (10, 20, 40 und 80 Items). In allen drei Studien wurde der Stichprobenumfang in drei Schritten variiert (N=100, 800, 6400). Dabei wurde versucht, das gesamte Spektrum von Stichprobengrößen abzubilden, das in der Praxis der Testkonstruktion vorkommt. Auch die Kommunalität bzw. Reliabilität des Kriteriums wurde in drei Schritten variiert (.3, .6, .9). Bei Ladung des Kriteriums auf mehreren gemeinsamen Faktoren ist immer eine Rotation im Raum der gemeinsamen Faktoren möglich, die dazu führt, dass das Kriterium schließlich nur noch auf einem Faktor lädt. Da die Rotation der Faktoren nichts an der resultierenden Kovarianzmatrix ändert, konnte die Ladung des Kriteriums auf einen Faktor begrenzt werden, ohne dass dadurch Einschränkungen der Allgemeinheit zu befürchten waren.

Die bisher besprochenen experimentellen Faktoren variierten zwischen den verschiedenen simulierten Itempools[11] (between-test factors). Das Selektionsverfahren und die Anzahl der ausgewählten Items sind dagegen Messwiederholungsfaktoren, die innerhalb eines Itempools variieren (within-test factors). Da der Umfang des Itempools die Anzahl der wählbaren Items begrenzt, konnten diese beiden Faktoren nicht vollständig miteinander gekreuzt werden. Stattdessen wurden die Ergebnisse für alle vier Abstufungen im Umfang des Itempools getrennt ausgewertet. Der Versuchsplan ist in Tabelle 4 (S. 69) skizziert.

[11] Die Anzahl der simulierten Personen ließe sich durchaus innerhalb der Tests variieren. Dies hätte jedoch den Nachteil, dass aus der Menge der Kovarianzmatrizen nur eine kleinere Stichprobe gezogen worden wäre, was eine Generalisierung der Ergebnisse einschränken würde.

Tabelle 4: Versuchsplan

Faktor	Stufen	Bemerkungen
Umfang des Itempools ***n***	10, 20, 40, 80	Getrennte Auswertung der Stufen
Anzahl der Faktoren *f*	2, 4, 6	Nur in Simulationsstudie 3 variiert
Kommunalität des Kriteriums	.3, .6, .9	In Simulationsstudie 1 nicht variiert, sondern gleich 1
Stichprobenumfang	100, 800, 6400	
Länge des Subtests	1,2, ... ,***n***	Messwiederholungsfaktor
Selektionsverfahren	• Vollständige Permutation • MAXVAL • Optimiertes Maxval • Trennschärfe • Itemvalidität • Cronbachs α	Messwiederholungsfaktor Vollständige Permutation nur bei Itempool mit 10 Items.

Für jede Zelle des Versuchsplans sollten mindestens 10 Tests, d.h. Kovarianzmatrizen, simuliert werden. In Studie 3a, 3b und 3c ergeben sich 108 (4x3x3x3) mögliche Kombinationen der Faktoren, die zwischen den Tests variieren. Daher wurden in jeder Studie 1080 Kovarianzmatrizen generiert. Da der Versuchsplan in Studie 1 und 2 nur 12 bzw. 36 Zellen hat, kommen hier nicht nur 10 Kovarianzmatrizen in jeder Zelle des Versuchsplans vor, sondern 90 in Studie 1 bzw. 30 in Studie 2. Aus jedem der simulierten Itempools (Kovarianzmatrizen) wurden für jede Testlänge diejenigen Itemteilmengen ermittelt, die nach Maßgabe der verschiedenen Selektionsalgorithmen am besten abschnitten. Alle untersuchten Selektionsalgorithmen griffen bei der Itemauswahl lediglich auf die Stichprobenkovarianzmatrix zurück. Die Validität der ausgewählten Skalen wurde jeweils anhand der Populationskovarianzmatrix bestimmt und verglichen. Anhand der Ladungsmatrix lässt auch die Reliabilität der Skalen in der Population exakt bestimmen, sofern man voraussetzt, dass die spezifische Varianz der Items ausschließlich unsystematische Fehlervarianz ist und keine wahre Varianz im Sinne der klassischen Testtheorie enthält. In Studie 1 macht dies jedoch keinen Sinn, da die Faktoren hier als Hauptkomponenten und nicht als latente Persönlichkeitsfaktoren zu interpretieren sind.

5.1.4 Zusammenfassende Beschreibung des Versuchsablaufs

Die Simulationsstudie gliedert sich in fünf verschiedene Teilstudien (Studie 1, 2, 3a, 3b, 3c), die getrennt ausgewertet werden, da jeweils ein anderes Modell für die Verteilung der Varianzkomponenten der Items zugrunde gelegt wurde (vgl. Tabelle 3 auf S. 66). Der Ablauf der Studien ist immer gleich (siehe auch Tabelle 2 auf S. 57):

1. In jeder dieser Studien wird eine Vielzahl von Faktorladungsmatrizen (siehe [5.1-13] auf S. 62) generiert. Die Ladungen der Items auf dem validen Faktor und auf den gemeinsamen Fehlerfaktoren werden per Zufallsgenerator bestimmt, während die Ladung auf dem spezifischen Fehlerfaktor fix ist (vgl. Tabelle 3 auf S. 66). Die Parameter werden dabei so gesetzt, dass alle bei dem jeweils zugrundegelegten Modell möglichen Datenstrukturen resultieren können und sich (im Durchschnitt) die gewünschte Verteilung der Varianzkomponenten ergibt (vgl. Tabelle 3 auf S. 66). Die Ladungen des Kriteriums auf dem Fehlerfaktor und dem validen Faktor werden so gesetzt, dass sich genau die gewünschte Kommunalität des Kriteriums ergibt (vgl. Tabelle 4 auf S. 69).

   Unterschiede zwischen den simulierten Faktorladungsmatrizen innerhalb einer Teilstudie ergeben sich nicht nur dadurch, dass einzelne Einträge per Zufallsgenerator bestimmt werden, sondern auch dadurch, dass die Struktur der gesamten Ladungsmatrix gezielt verändert wird (vgl. Tabelle 3 auf S. 66 und Tabelle 4, S. 69). Variationen in der Struktur der Ladungsmatrix betreffen den Umfang des Itempools, die Anzahl der gemeinsamen Faktoren und die Kommunalität des Kriteriums (vgl. Tabelle 4 auf S. 69).

2. Zu jeder Faktorladungsmatrix wird jeweils eine Personenstichprobe gezogen. Die Faktorwerte der Versuchspersonen werden durch zufällige Ziehung aus der Standardnormalverteilung bestimmt. Die Anzahl der Versuchspersonen variiert *zwischen* den Faktorladungsmatrizen (vgl. Tabelle 4 auf S. 69).

3. Durch die Multiplikation der Faktorladungsmatrix mit der Matrix der Faktorwerte der Versuchspersonen ergibt sich die Matrix der Itemscores der Versuchspersonen. Aus dieser lässt sich die Stichprobenkovarianzmatrix bestimmen. Die Stichprobenkovarianzmatrix ist die Grundlage für die verschiedenen Itemselektionsmethoden (vgl. Tabelle 1 auf S. 49 und Tabelle 4 auf S. 69). Anhand der Stichprobenkovarianzmatrix lässt sich für jede ausgewählte Itemkombination auch in der Stichprobe die externe Validität bestimmen.

4. Durch Multiplikation der Faktorladungsmatrix mit ihrer Transponierten bestimmt man die Kovarianzmatrix in der Population (vgl. [5.1-11] auf S. 62). Anhand der Faktorladungsmatrix lässt sich auch die Reliabilität bzw. Kommunalität der Items und des Kriteriums bestimmen. Anhand der Populationskovarianzmatrix sowie der Reliabilität der Items und des Kriteriums lässt sich für jede ausgewählte Itemkombination sowohl die Validität (vgl. [2.2-2] auf S. 31) und die Reliabilität (vgl. Krauth, 1995, S. 269) als auch das in Kapitel 3.2 definierte Suppressionskriterium ([3.2-5] auf S. 44) berechnen.

5.1.5 Statistische Auswertung

Auf den ersten Blick bietet sich bei dem vorliegenden Versuchsplan (siehe Tabelle 4 auf S. 69) eine Varianzanalyse mit Messwiederholung als statistisches Verfahren zur Auswertung der Daten an. Bei der üblichen univariaten Auswertung von Versuchsplänen mit Messwiederholung muss man jedoch voraussetzen, dass die Kovarianzmatrizen der abhängigen Variablen die Sphärizitäts- bzw. Zirkularitätsannahme erfüllen (Werner, 1997). Bei der vorliegenden Abhängigkeitsstruktur der Daten dürfte diese Annahme kaum gegeben sein, da die Varianz der Differenzen zwischen Skalen verschiedener Länge umso größer sein dürfte, je größer der Unterschied in der Skalenlänge ist. Mit Hilfe gemischter Modelle (Littell, Milliken, Stroup & Wolfinger, 1996) lassen sich auch andere Abhängigkeitsmuster als die Sphärizität modellieren (z.B. autoregressive Fehler). Wenngleich bei der vorliegenden Studie durchaus begründete Annahmen über die Struktur der Kovarianzmatrix der abhängigen Variablen gemacht werden können, sind diese jedoch nicht so präzise, dass vorab eine bestimmte Abhängigkeitsstruktur der Daten zugrundegelegt werden konnte. In diesem Fall ist eine multivariate Auswertung der Daten angezeigt (O'Brian & Kaiser, 1985). Dazu werden die verschiedenen Stufen der Messwiederholungsfaktoren als abhängige Variable betrachtet. Hierbei werden keine Annahmen über die Abhängigkeitsstruktur der Daten gemacht. Voraussetzung für dieses Verfahren ist lediglich, dass die abhängigen Variablen multivariat normalverteilt sind und dass deren Kovarianzmatrix in allen Zellen des Versuchplans gleich ist. Als Teststatistik wurde Wilks Lambda verwendet (Stevens, 1979; Olson, 1976).

Bei Messwiederholungsdesigns kann man sich bei der Durchführung von Poweranalysen nicht auf etablierte Konventionen zur Beurteilung von Effektstärken stützen, wie sie beispielsweise von Cohen (1988) für zahlreiche univariate statistische Verfahren publiziert wurden. Dass die Effektstärke bei multivariaten Verfahren nicht nur von Ausmaß der Mittelwertsunterschiede, sondern auch in erheblichen Maß von den Korrelation der abhängigen Variablen abhängt,

erschwert die a-priori Abschätzung der Teststärke (Muller, LaVange, Ramey & Ramey, 1992; Stevens, 1980). Daher wurde von der Durchführung von Poweranalysen abgesehen.

5.2 Ergebnisse

Grundsätzlich ist bei allen durchgeführten Simulationsstudien nur der Effekt der Selektionsmethode sowie dessen Interaktion mit den anderen Effekten interessant. Daher wird auf die Darstellung der anderen Effekte weitgehend verzichtet. Um bei der Vielzahl der durchgeführten Signifikanztests die Gefahr der Interpretation von Zufallseffekten relativ gering zu halten, werden nur solche Effekte interpretiert, die auf dem 1-Promille Niveau signifikant sind.

Die Unterschiede zwischen dem MAXVAL-Verfahren und dessen Optimierung durch Vertauschungen einzelner Items sowie der vollständigen Permutation der Items waren so gering, dass auf eine getrennte Darstellung verzichtet wird. Stattdessen werden nur die Ergebnisse des MAXVAL-Verfahrens berichtet.

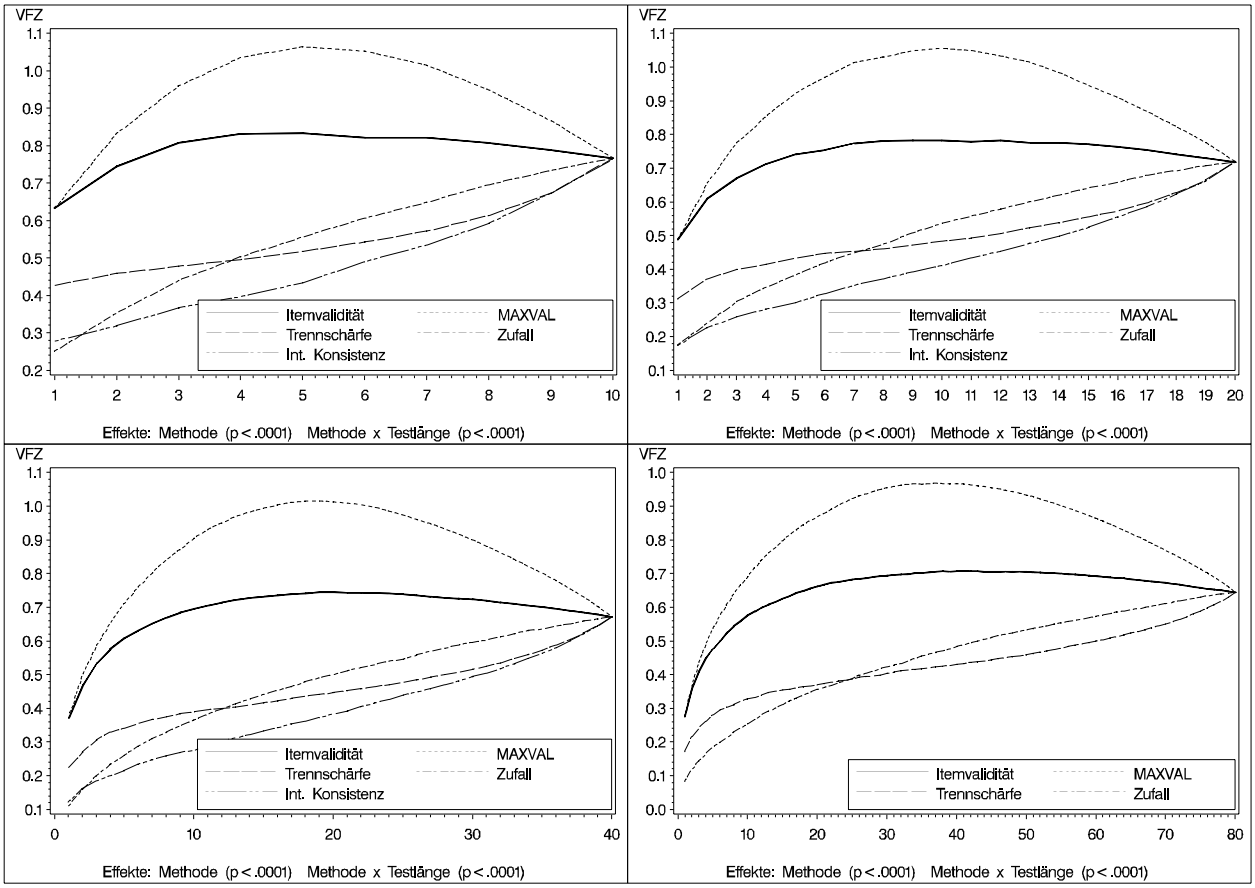
Da in der Simulationsstudie mehrere tausend Mittelwerte verglichen wurden, werden die Ergebnisse nicht numerisch, sondern ausschließlich in Form von Graphiken präsentiert. Die Graphiken stellen entweder den Haupteffekt der Selektionsmethode sowie dessen Interaktion mit der Testlänge dar oder auch höhere Interaktionen dieser beiden Effekte mit anderen Variablen. In den Graphiken werden auf der Ordinate jeweils die Mittelwerte der Reliabilität (REL), der Fisher-Z standardisierten Validität (VFZ) oder des numerischen Maßes für Suppressionseffekte (SUP) dargestellt, während auf der Abszisse die Testlänge angegeben ist. Der Umfang des Itempools entspricht jeweils dem maximalen Wert der Testlänge auf der Abszisse. Die Ergebnisse der verschiedenen Selektionsverfahren werden jeweils innerhalb einer Graphik differenziert dargestellt. Sofern die Ausprägung der anderen Faktoren nicht in der Titelzeile der Graphik aufgeführt sind, wurden die Werte über alle Stufen des jeweiligen Faktors aggregiert. In der Fußzeile finden sich jeweils Angaben zur Signifikanz der dargestellten Effekte (p-Werte). Falls kein Faktor in der Titelzeile der Graphik genannt wird, so beziehen sich die statistischen Tests auf den Haupteffekt der Selektionsmethode sowie dessen Interaktion mit der Testlänge. Interaktionseffekte der Selektionsmethode mit anderen Faktoren als der Testlänge können nicht in einer einzigen Graphik dargestellt werden. Sie ergeben sich aus dem Vergleich von mehreren Graphiken, bei denen die Ausprägungen der (anderen) beteiligten Faktoren in der Titelzeile der Graphik auftauchen. Die Fußzeilen von verschiedenen Graphiken, die denselben Interaktionseffekt darstellen, sind daher identisch; auch die Prädiktoren, die in der Titelzeile

vorkommen – nicht jedoch deren Ausprägungen – sind dann identisch. Die verschiedenen Graphiken geben also die Reliabilität bzw. Validität auf den jeweiligen Stufen der an dem Interaktionseffekt beteiligten Faktoren wieder.

5.2.1 Studie 1

Haupteffekt der Methode und deren Interaktion mit der Testlänge

Abbildung 2: Validität der verschiedenen Selektionsverfahren in Studie 1 (in der Population) für die vier verschiedenen Testlängen (10, 20, 40, 80)



In der ersten Simulationsstudie wurde versucht sicherzustellen, dass jede beliebige Populationskovarianzmatrix vorkommen kann (vgl. Phase 1 in Tabelle 2 auf S. 57 sowie Tabelle 3 auf S. 66). Hier war das MAXVAL-Verfahren den anderen Verfahren deutlich überlegen. Die aufgrund der Stichprobenkovarianzmatrizen zusammengestellten Skalen zeigten beim MAXVAL-Verfahren mit Abstand die höchste Validität in den zugrundeliegenden Populationen (vgl. Abbildung 2). Nur bei diesem Verfahren konnte durch Elimination von Items eine deutliche Steigerungen der Validität erreicht werden. Entfernt man dagegen schrittweise die Items mit der geringsten Validität, so konnten allenfalls leichte Steigerungen der Validität beobachtet werden. Aber auch bei Selektion anhand der Itemvalidität können die Skalen im Durchschnitt auf ein

Fünftel ihrer ursprünglichen Länge reduziert werden, ohne dass es zu deutlichen Einbußen bei der Validität kommt.

Wie erwartet, führt die Selektion anhand der Trennschärfe bei kurzen Skalen zu einer besseren Validität als bei zufälliger Auswahl. Die Entfernung von wenigen Items mit geringer Trennschärfe führt dagegen zu Skalen, die weniger valider sind als bei zufälliger Selektion. Es fällt auf, dass die Abnahme der Validität bei Selektion anhand der Trennschärfe nicht gleichmäßig beschleunigt ist. Nach einer relativ deutlichen Abnahme der Validität bei Eliminierung der Items mit der geringsten Trennschärfe stabilisiert sich die Validität zunächst ein wenig, um dann bei der Entfernung der trennschärfsten Items wieder deutlich abzufallen. Eine Maximierung der internen Konsistenz (Cronbachs Alpha) führt zu sehr ähnlichen Ergebnissen wie die Selektion anhand der Trennschärfe, wenn nur ein kleiner Teil der Items entfernt wird. Sonst führt die Selektion anhand der internen Konsistenz durchweg zu Skalen geringerer Validität als bei den anderen Verfahren einschließlich der zufälligen Auswahl. Die Ergebnisse in der Stichprobe unterscheiden sich nur unwesentlich von den Ergebnissen in der Population (vgl. Abbildung 3).

Abbildung 3: Validität der verschiedenen Selektionsverfahren in Studie 1 (in der Stichprobe)

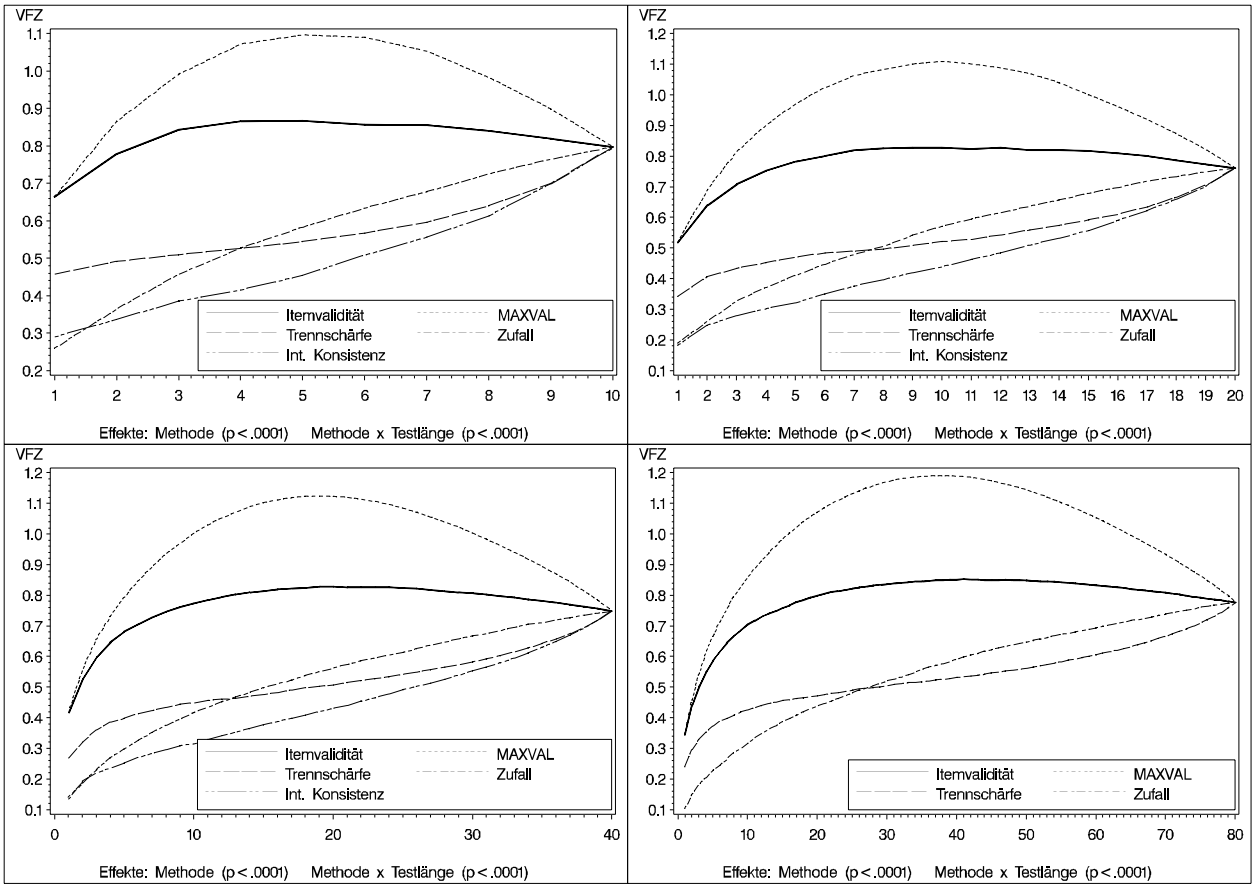


Abbildung 4: Interaktion des Stichprobenumfangs mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Population (Fisher Z-transformiert) in Studie 1 (getrennte Darstellung je nach Umfang des Itempools).

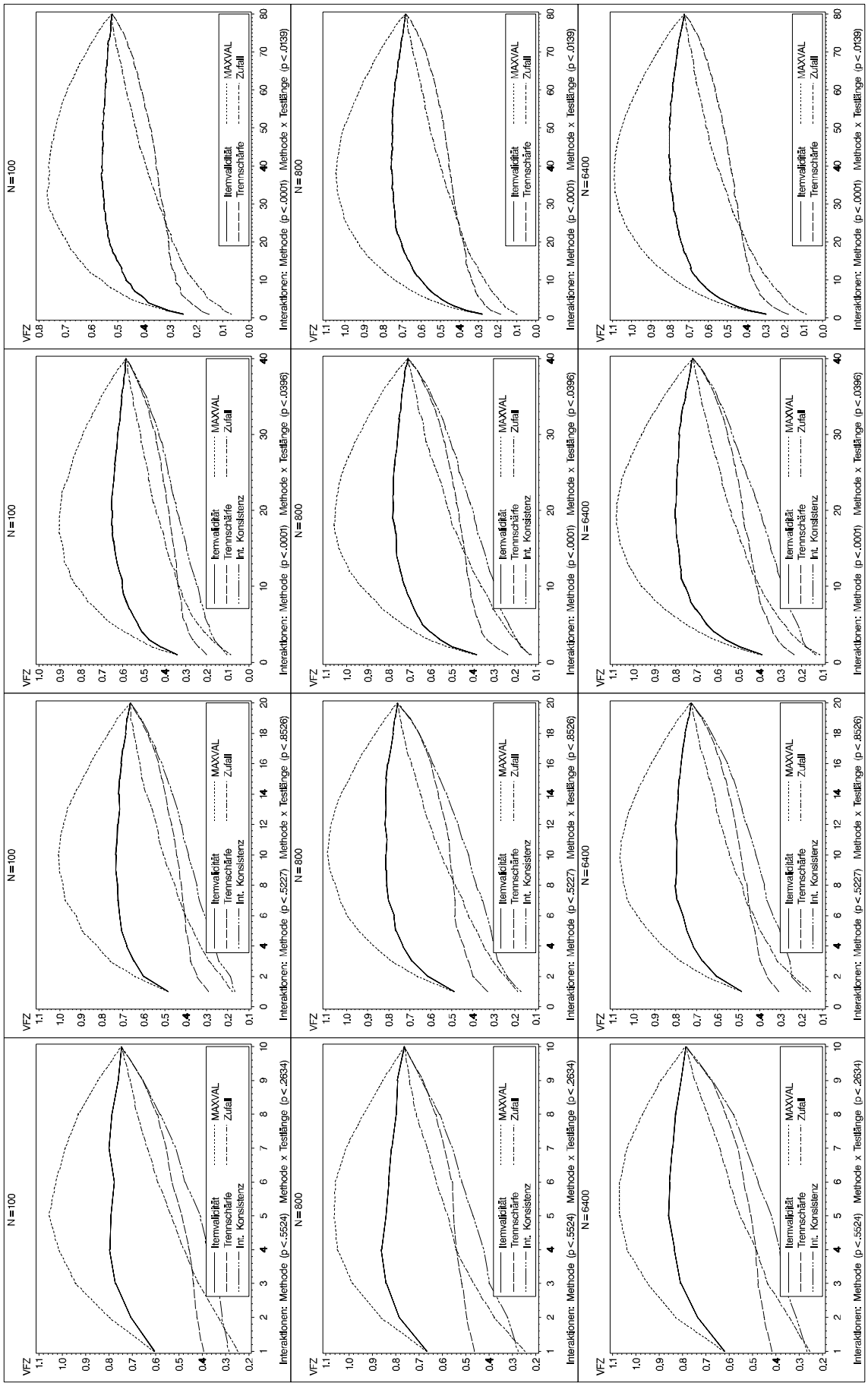
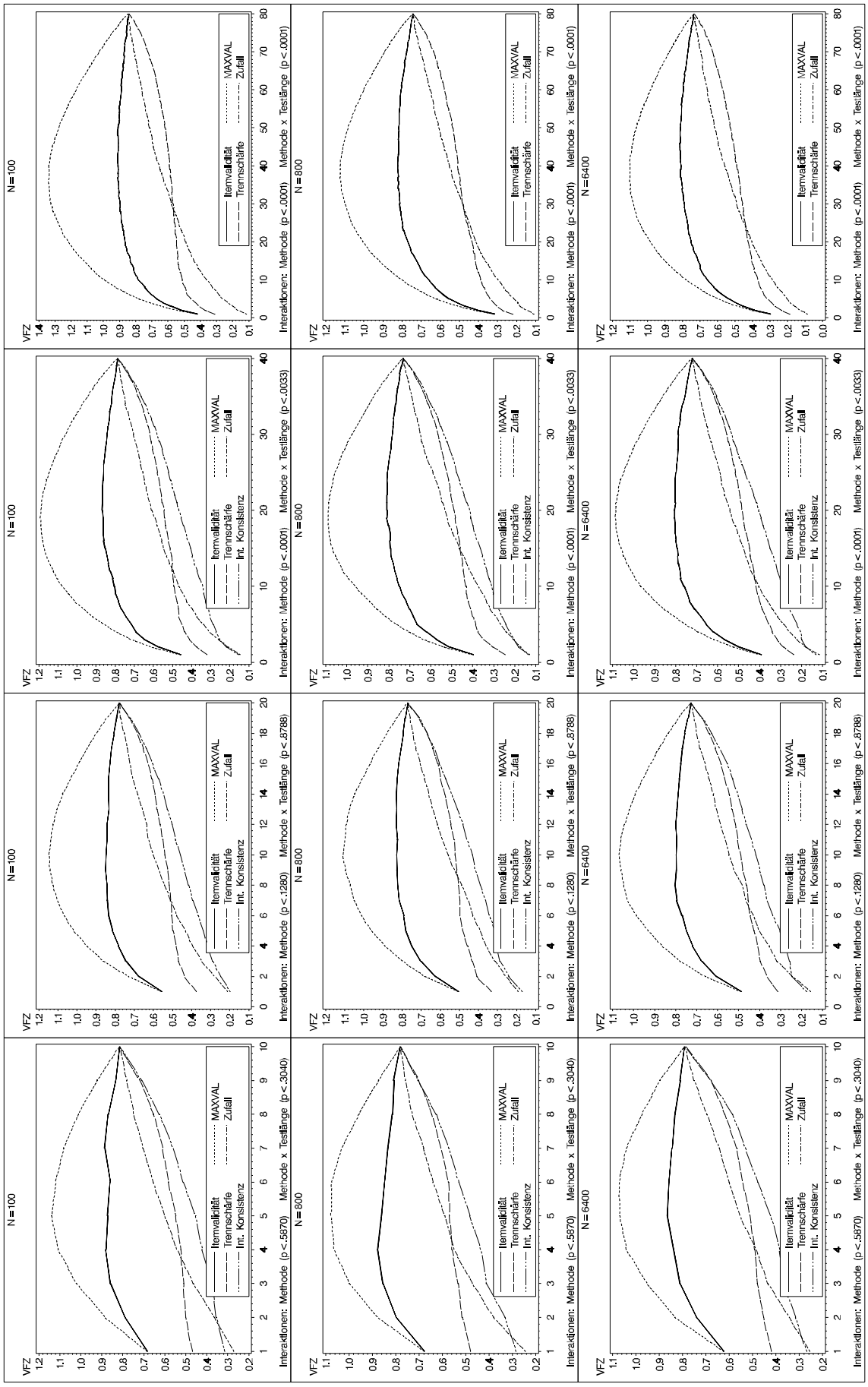


Abbildung 5: Interaktion des Stichprobenumfangs mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Stichprobe (Fisher Z-transformiert) in Studie 1 (getrennte Darstellung je nach Umfang des Itempools).



Interaktion zwischen der Selektionsmethode und dem Stichprobenumfang

Nur bei umfangreichen Itempools zeigte sich eine signifikante Interaktion des Stichprobenumfangs mit der Selektionsmethode. In der Population scheint diese Interaktion darauf zurückzuführen zu sein, dass sich die Selektionsverfahren mit zunehmendem Stichprobenumfang immer deutlicher unterscheiden (vgl. die Skalierung der letzten beiden Spalten in Abbildung 4 auf S. 75).

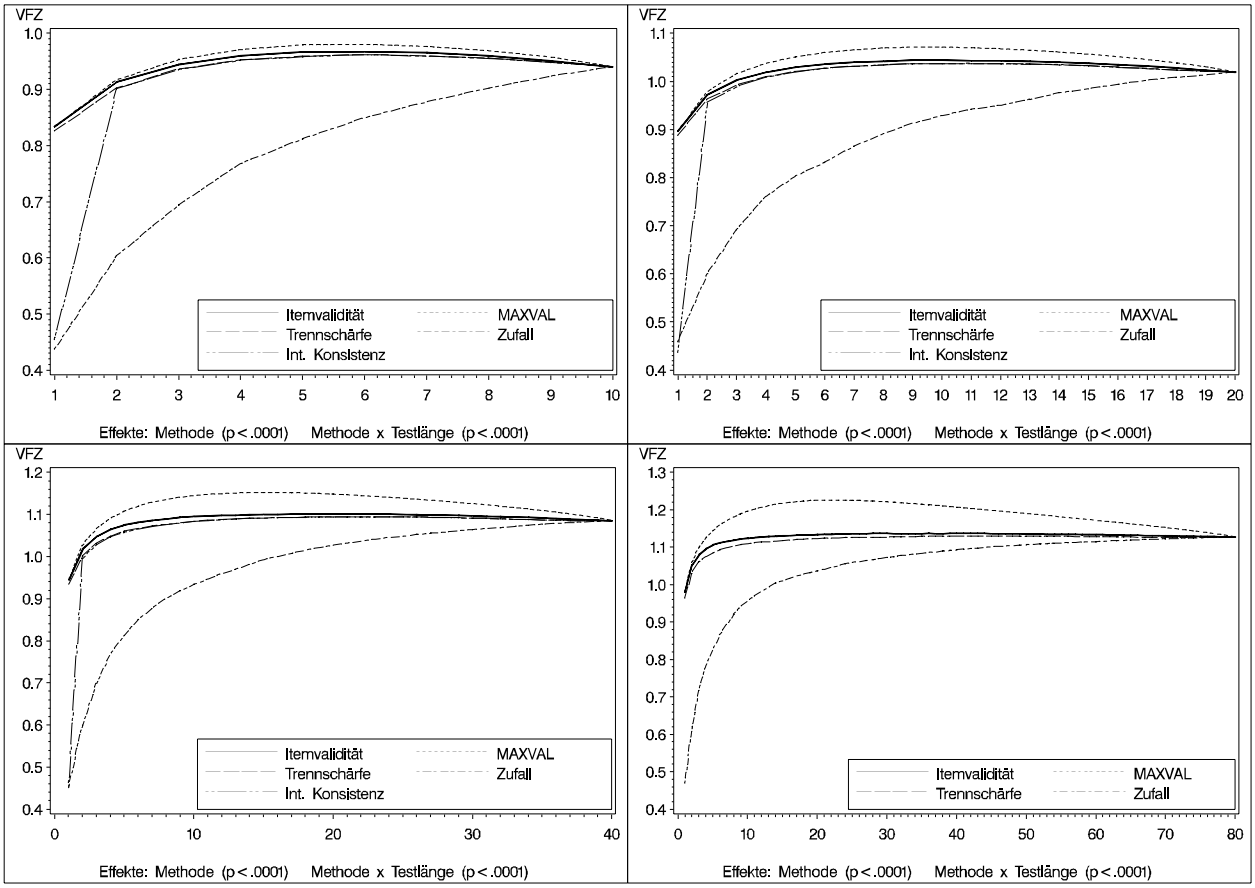
In der Stichprobe nehmen die Unterschiede zwischen den Selektionsverfahren mit zunehmendem Stichprobenumfang dagegen ab (vgl. die Skalierung der letzten beiden Spalten in Abbildung 5 auf S. 76).

5.2.2 Studie 2

Validität

Haupteffekt der Methode und deren Interaktion mit der Testlänge

Abbildung 6: Validität der verschiedenen Selektionsverfahren in Studie 2 (in der Stichprobe)



In der zweiten Simulationsstudie sollten die Selektionsmethoden bei einem eindimensionalen, τ-kongenerischen Itempool verglichen werden. Das MAXVAL-Verfahren erreichte hier nur bei umfangreichen Itempools in der Stichprobe deutlich höhere Validitäten als die anderen Verfahren, während die Selektion anhand der Itemvalidität nur unwesentlich besser abschnitt als die Selektion anhand der Trennschärfe und die Optimierung von Cronbachs α (vgl. Abbildung 6 auf 77).

In der Population stellen sich die Verhältnisse jedoch ganz anders dar. Hier waren die Selektion anhand der Trennschärfe oder der Itemvalidität sowie die Optimierung der internen Konsistenz diejenigen Verfahren, bei denen sich die höchste Validität ergab, während das MAXVAL-Verfahren vor allem bei umfangreichen Itempools deutlich schlechter abschnitt (vgl. Abbildung 7).

Abbildung 7: Validität der verschiedenen Selektionsverfahren in Studie 2 (in der Population)

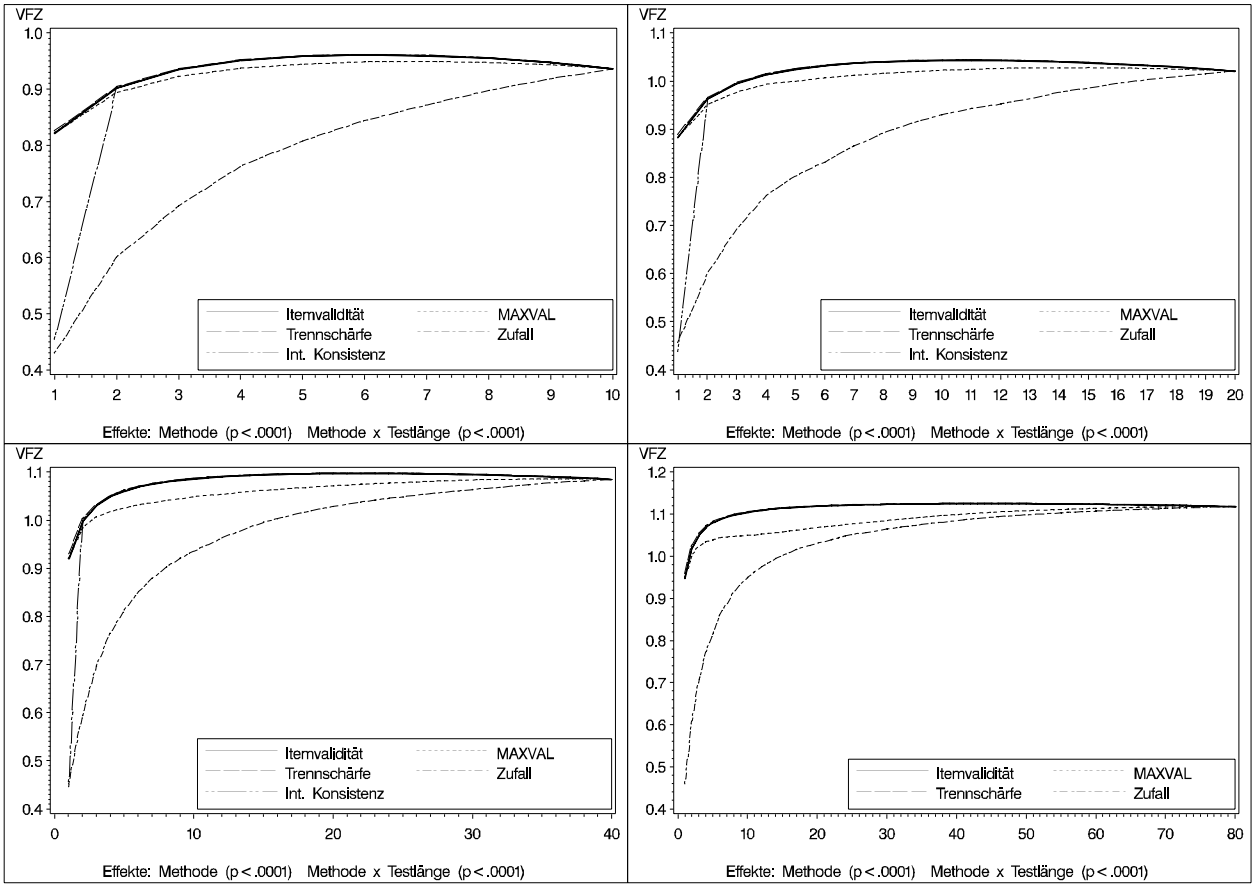


Abbildung 8: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Validität in der Stichprobe in Studie 2 (Testlänge 10)

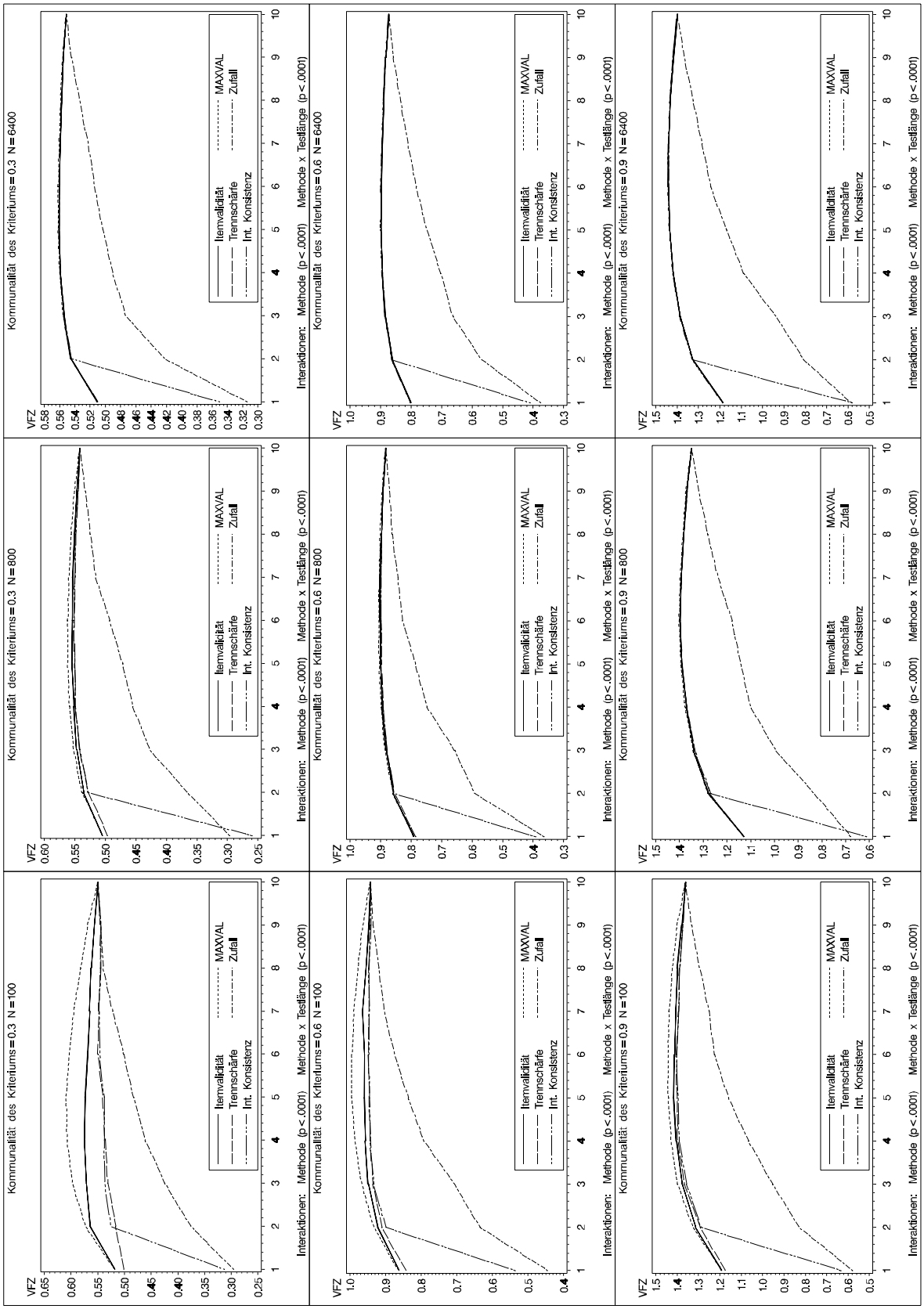


Abbildung 9: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Validität in der Population in Studie 2 (Testlänge 10)

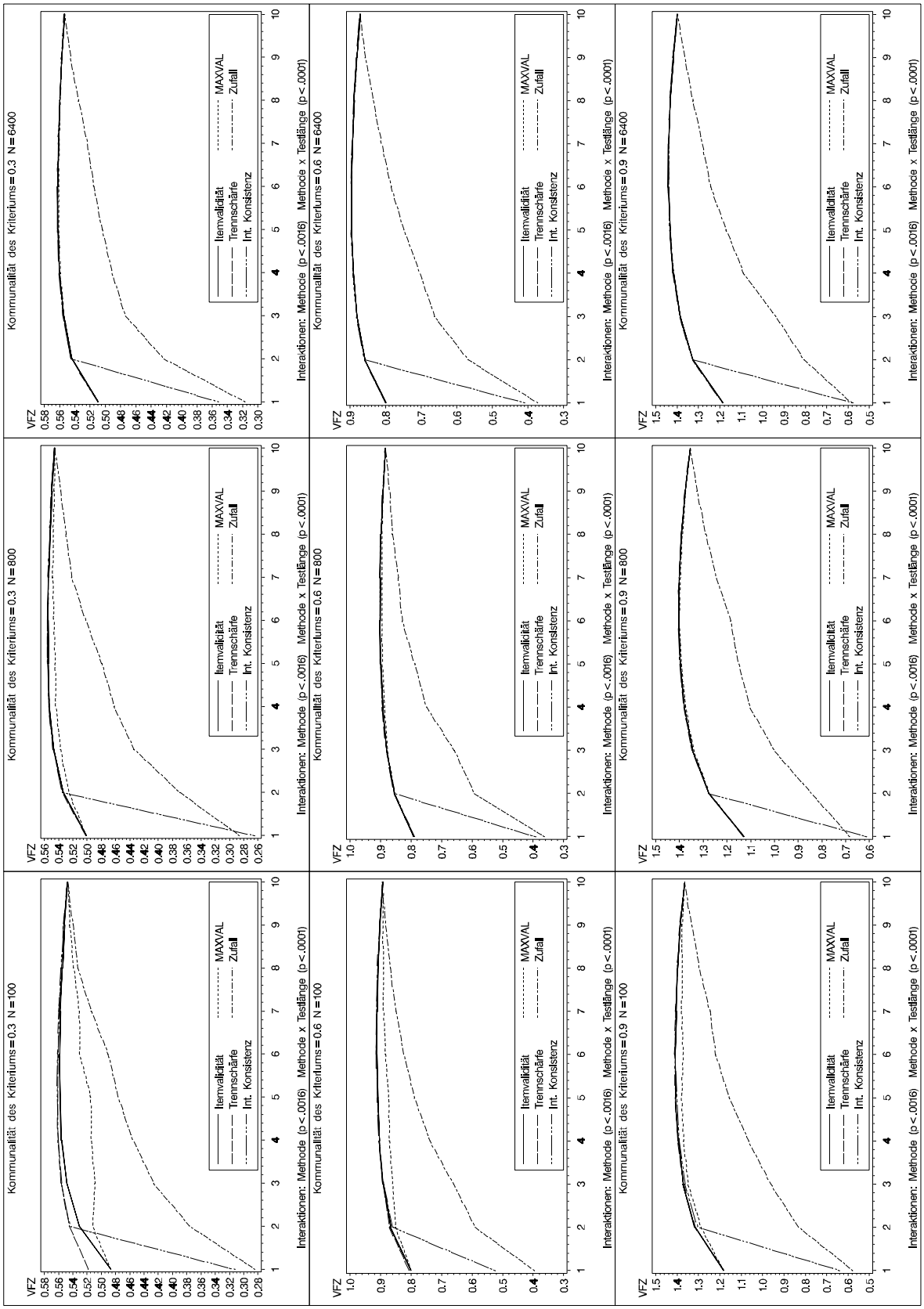


Abbildung 10: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Validität in der Stichprobe in Studie 2 (Testlänge 20)

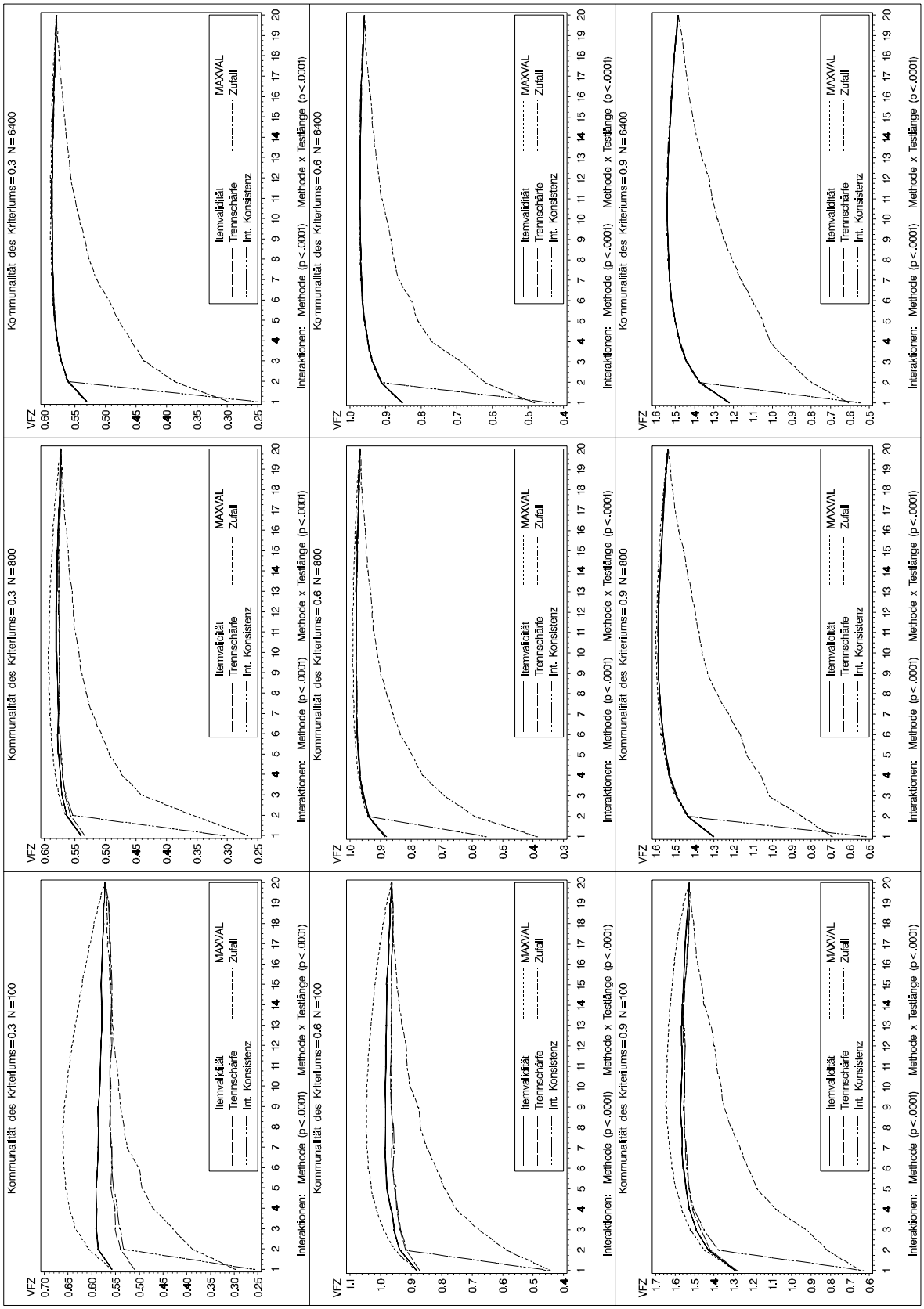


Abbildung 11: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Validität in der Population in Studie 2 (Testlänge 20)

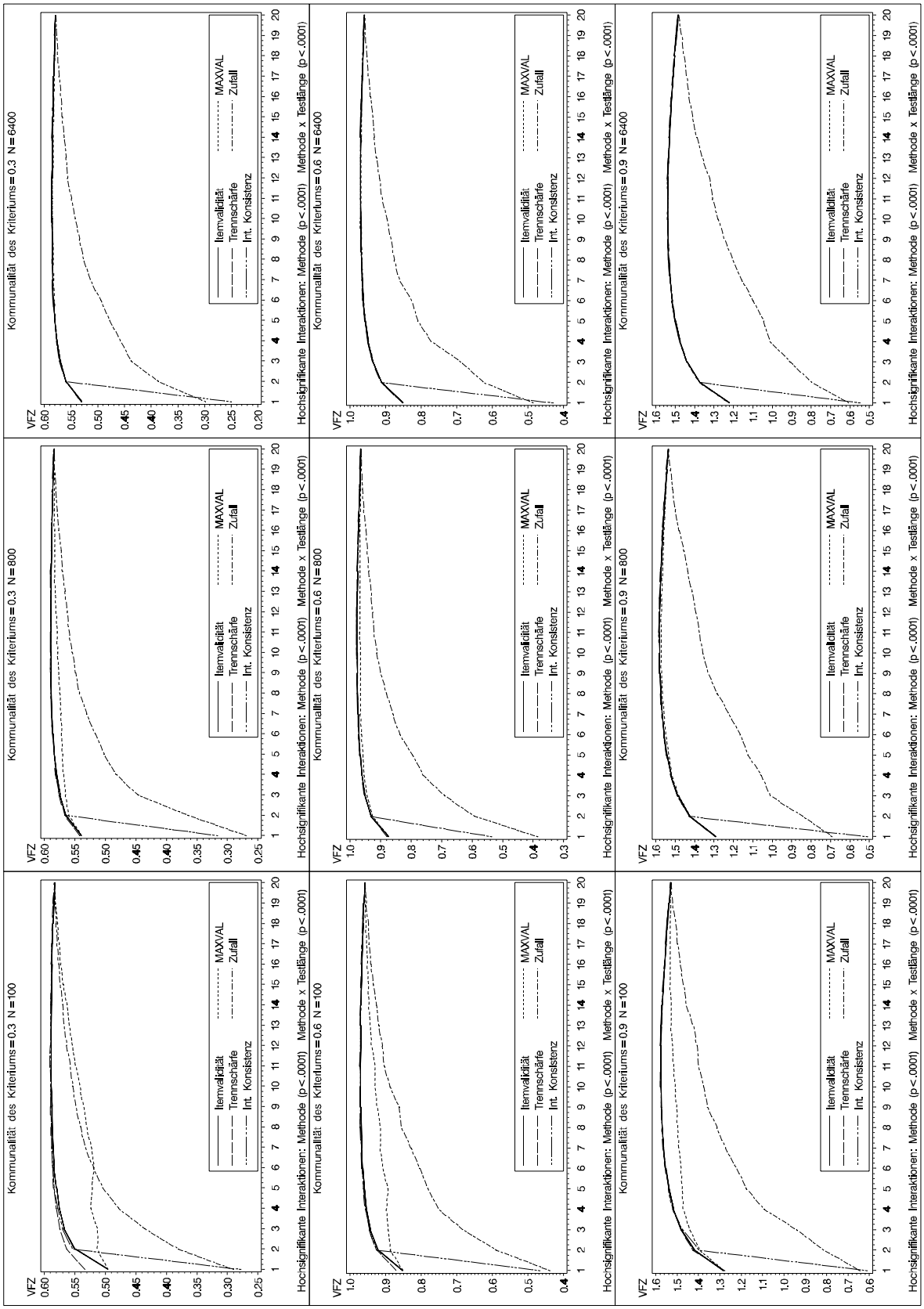


Abbildung 12: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Validität in der Stichprobe in Studie 2 (Testlänge 40)

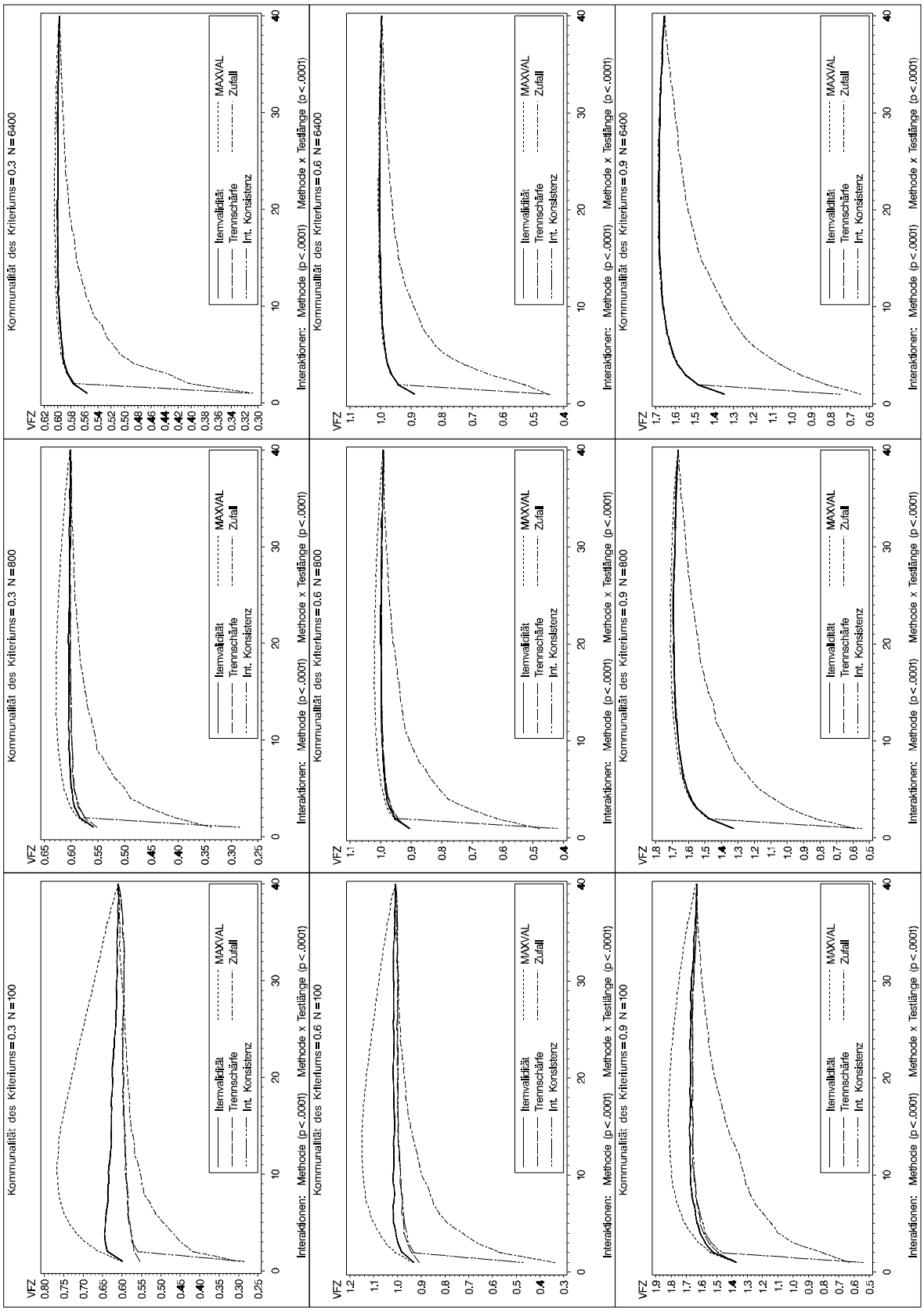


Abbildung 13: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Validität in der Population in Studie 2 (Testlänge 40)

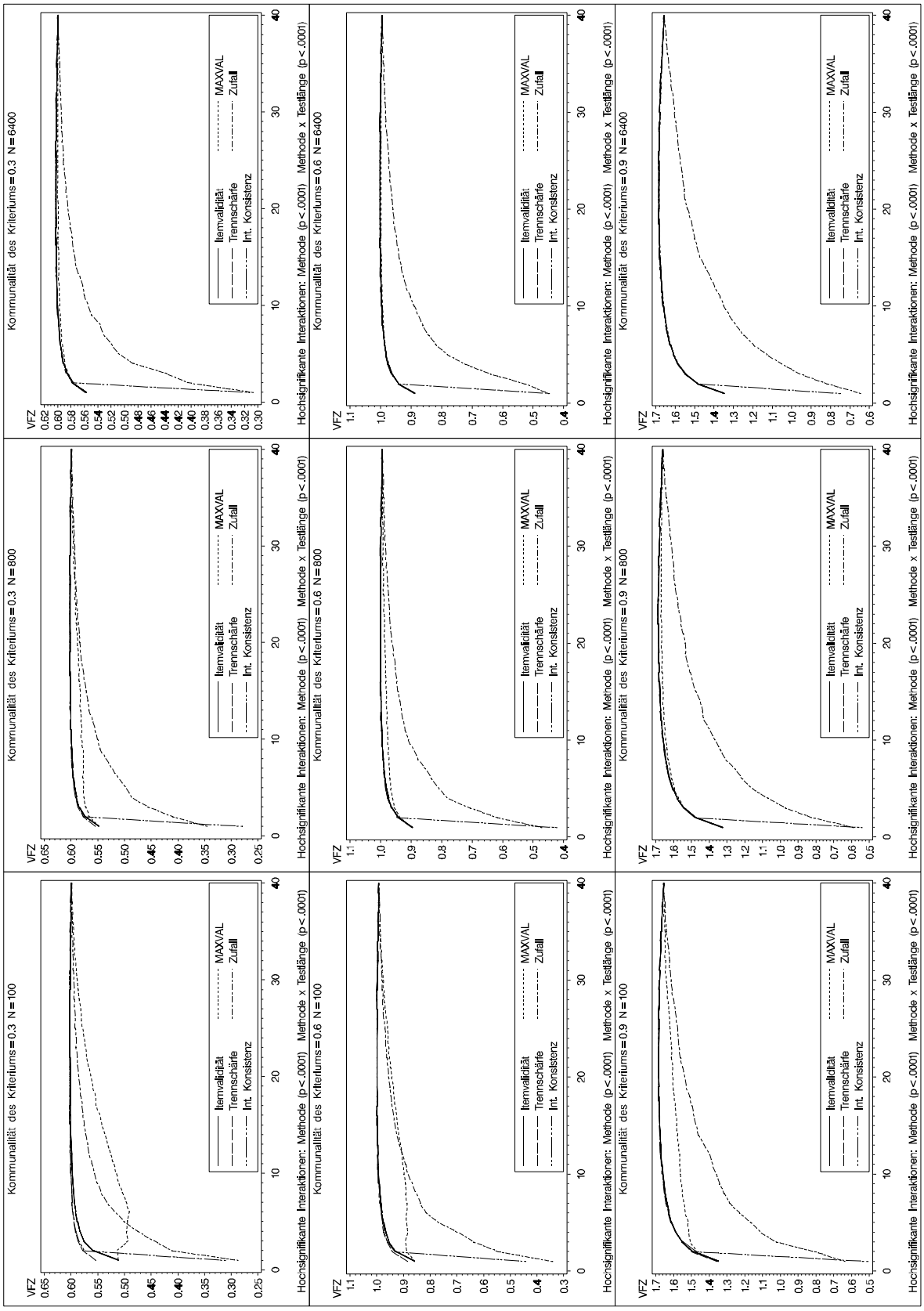


Abbildung 14: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Validität in der Stichprobe in Studie 2 (Testlänge 80)

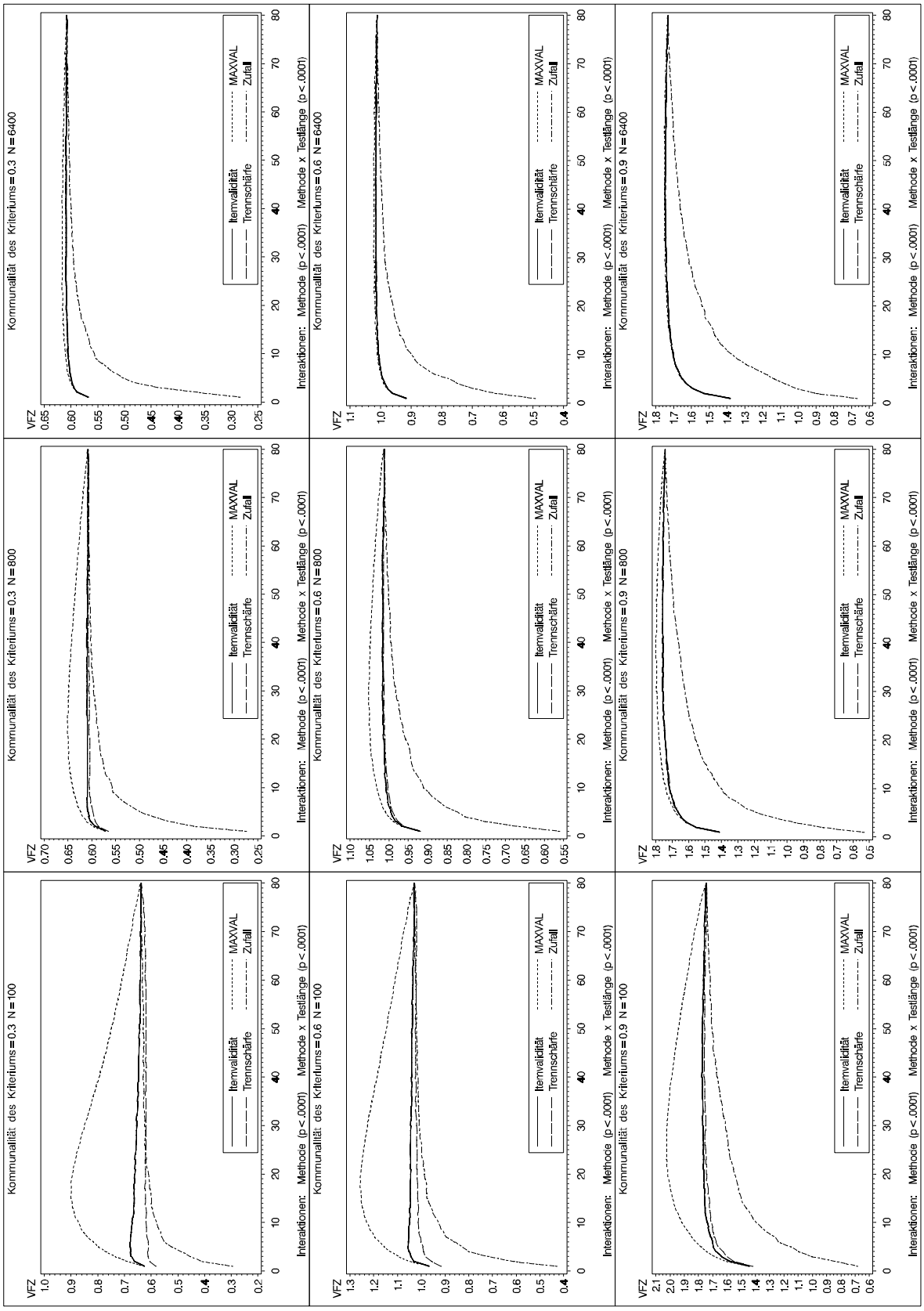
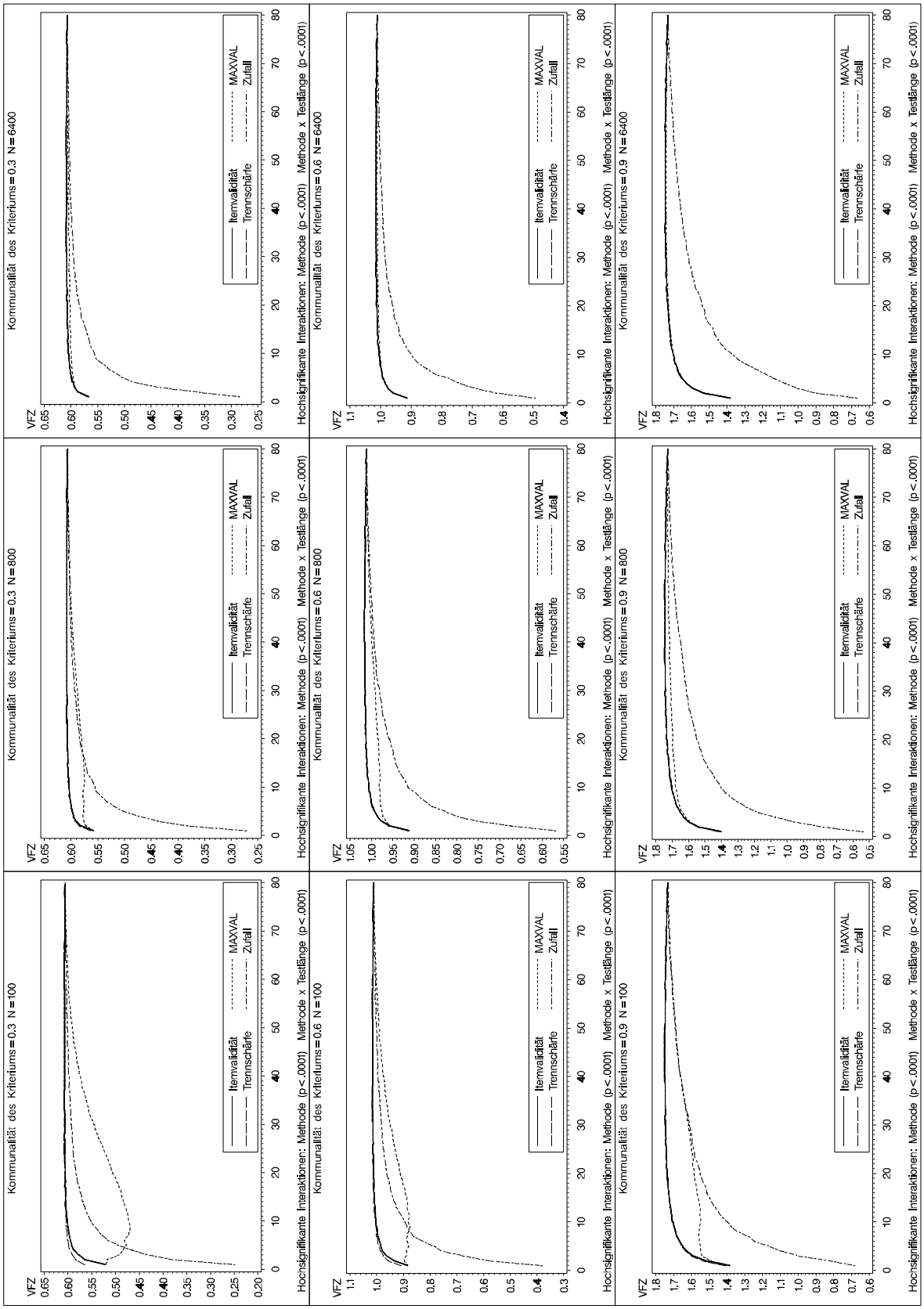


Abbildung 15: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Validität in der Population in Studie 2 (Testlänge 80)



Interaktion zwischen der Selektionsmethode, dem Stichprobenumfang und der Kommunalität des Kriteriums

Bei der Analyse der Interaktion zwischen der Selektionsmethode, dem Stichprobenumfang und der Kommunalität des Kriteriums[12] fällt auf, dass das MAXVAL-Verfahren bei großen Stichprobenumfängen zu sehr ähnlichen Resultaten führt wie die anderen Verfahren (vgl. Abbildung 8 auf S. 79 bis Abbildung 15 auf S. 86). Auch bei kleinen Stichprobenumfängen schneidet das MAXVAL-Verfahren nicht wesentlich schlechter ab als die anderen Verfahren, wenn die Kommunalität des Kriteriums hoch ist. Bei geringer Kommunalität des Kriteriums und kleinen Stichproben ist das MAXVAL-Verfahren jedoch – v.a. bei größeren Itempools – in der Population sogar deutlich schlechter als die zufällige Auswahl, während in der Stichprobe eine gravierenden Überschätzung der Validität zu beobachten ist. In diesem Fall wird die Validität auch bei Selektion anhand der Itemvalidität in der Stichprobe überschätzt, während in der Population bei sehr kurzen Skalen etwas weniger valide Tests resultieren als bei Selektion anhand der Trennschärfe oder der Optimierung der internen Konsistenz.

Reliabilität

Bei einem τ-kongenerischen Itempool ist eine höhere Validität zwangsläufig mit einer höheren Reliabilität verbunden, sofern die Fehler der Items nicht mit dem Fehler des Kriteriums korrelieren (vgl. Kapitel 1.3, S. 22). Da zwischen der Validität und der Reliabilität jedoch lediglich ein nicht-linearer (monotoner) Zusammenhang besteht, muss der Vergleich der Selektionsverfahren nicht zwangsläufig zu denselben Ergebnissen führen, wenn die Mittelwerte von mehreren Itempools miteinander verglichen werden. In der vorliegenden Studie ergaben sich jedoch keine nennenswerten Unterschiede zwischen den Ergebnissen der Reliabilität und denen in Bezug auf die Validität, so dass auf die Darstellung der Reliabilitätsergebnisse verzichtet wird.

Suppression

In Studie 2 sind die Items sowohl untereinander als auch hinsichtlich des Kriteriums τ-kongenerisch. Unter diesen Bedingungen kann in der Population weder Fehlersuppression noch Fehlerredundanz vorliegen (vgl. [3.2-8] auf S. 46 und [1.1-17] auf S. 10). Schätzt man jedoch bei bekannter Kommunalität (bzw. Reliabilität) des Kriterium mithilfe von [3.2-4] (S. 44) anhand der Stichprobendaten die Korrelation der Residuen der Items bei Herauspartialisierung

[12] Die 2-Weg Interaktionen der Selektionsmethode mit der Testlänge und der Kommunalität des Kriteriums sind ebenso wie die 3-Weg Interaktion unter Beteiligung der Testlänge sowohl in der Population als auch in der Stichprobe signifikant ($p<.0001$). Von Darstellung der entsprechenden aggregierten Mittelwerte der Validität wurde jedoch abgesehen.

der wahren Kriteriumswerte, so ergeben sich durchaus von null verschiedene Werte (vgl. Abbildung 16 auf S. 88). Dies lässt sich auf Stichprobenfehlern bei der Schätzung der Kovarianzmatrix zurückführen. Abbildung 16 (S. 88) bis Abbildung 18 (S. 90) geben die Stichprobenkennwerteverteilung des vorgestellten Suppressionskriterium bei Geltung der Nullhypothese (=bei keiner Itemkombination liegt Fehlersuppression oder Fehlerredundanz vor) wieder. An der Skalierung der Ordinaten in Abbildung 16 (S. 88) bis Abbildung 18 (S. 90) erkennt man, dass das MAXVAL-Verfahren v.a. bei geringer Kommunalität des Kriteriums und geringem Stichprobenumfang versucht von vermeintlichen Suppressionseffekten zu profitieren, während man bei Selektion anhand der Trennschärfe oder der internen Konsistenz anhand der Stichprobendaten mitunter fälschlicherweise den Eindruck gewinnen könnte, dass Fehlerredundanz gegeben ist.

Abbildung 16: Kovarianz der Residuen bei Herauspartialisierung der wahren Kriteriumswerte

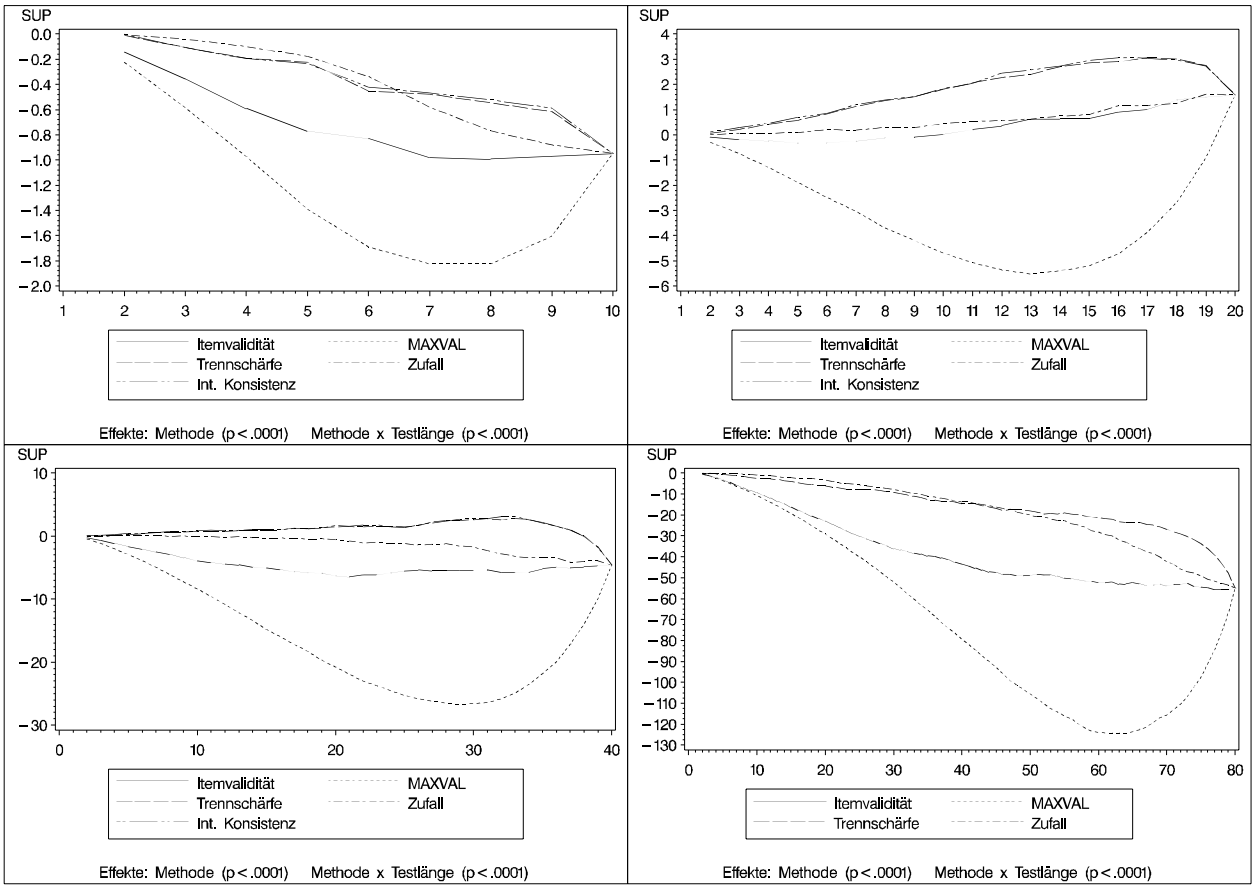


Abbildung 17: Interaktion des Stichprobenumfangs mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Kovarianz der Residuen in der Stichprobe (Fisher Z-transformiert) in Studie 2 (getrennte Darstellung je nach Umfang des Itempools).

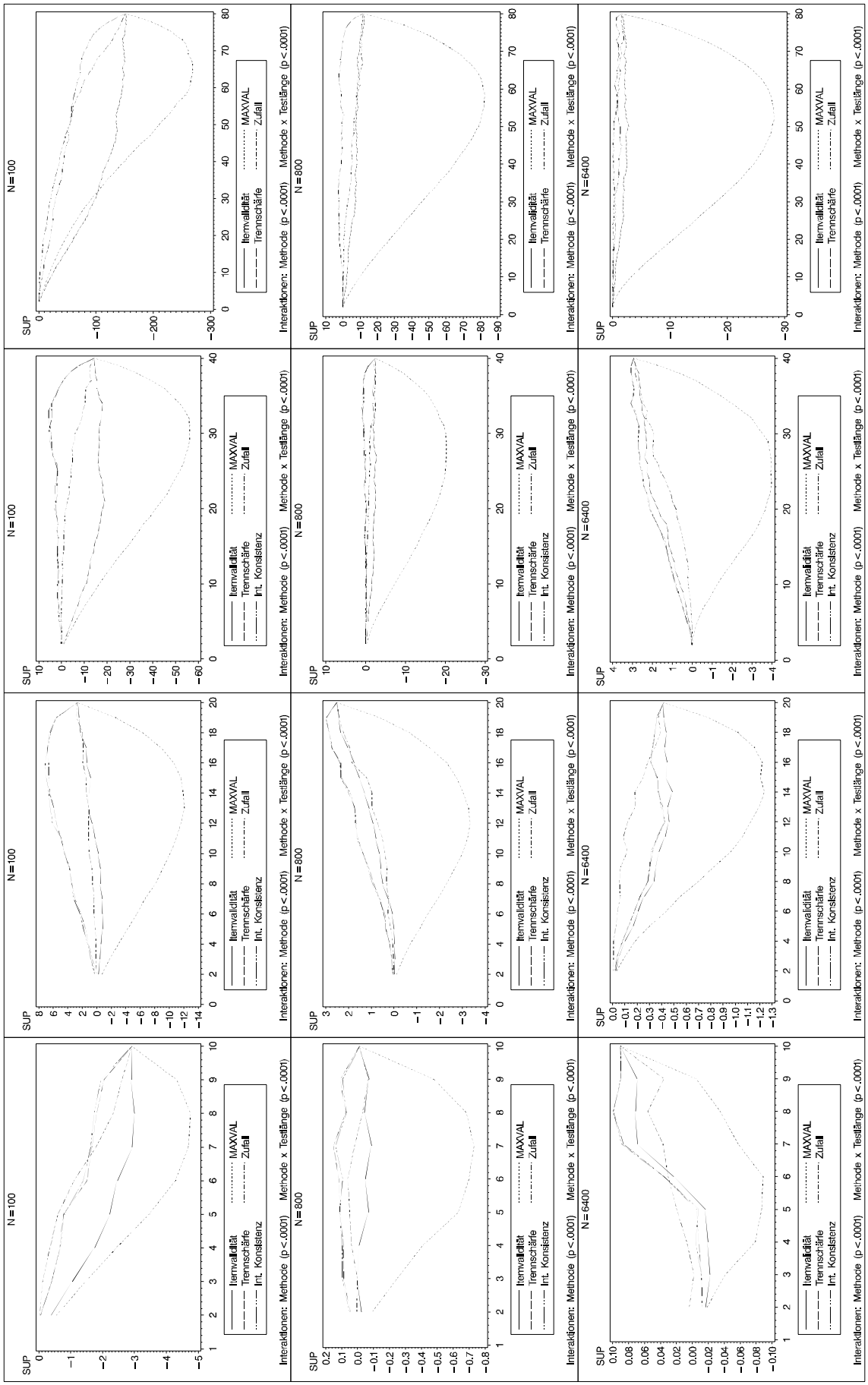
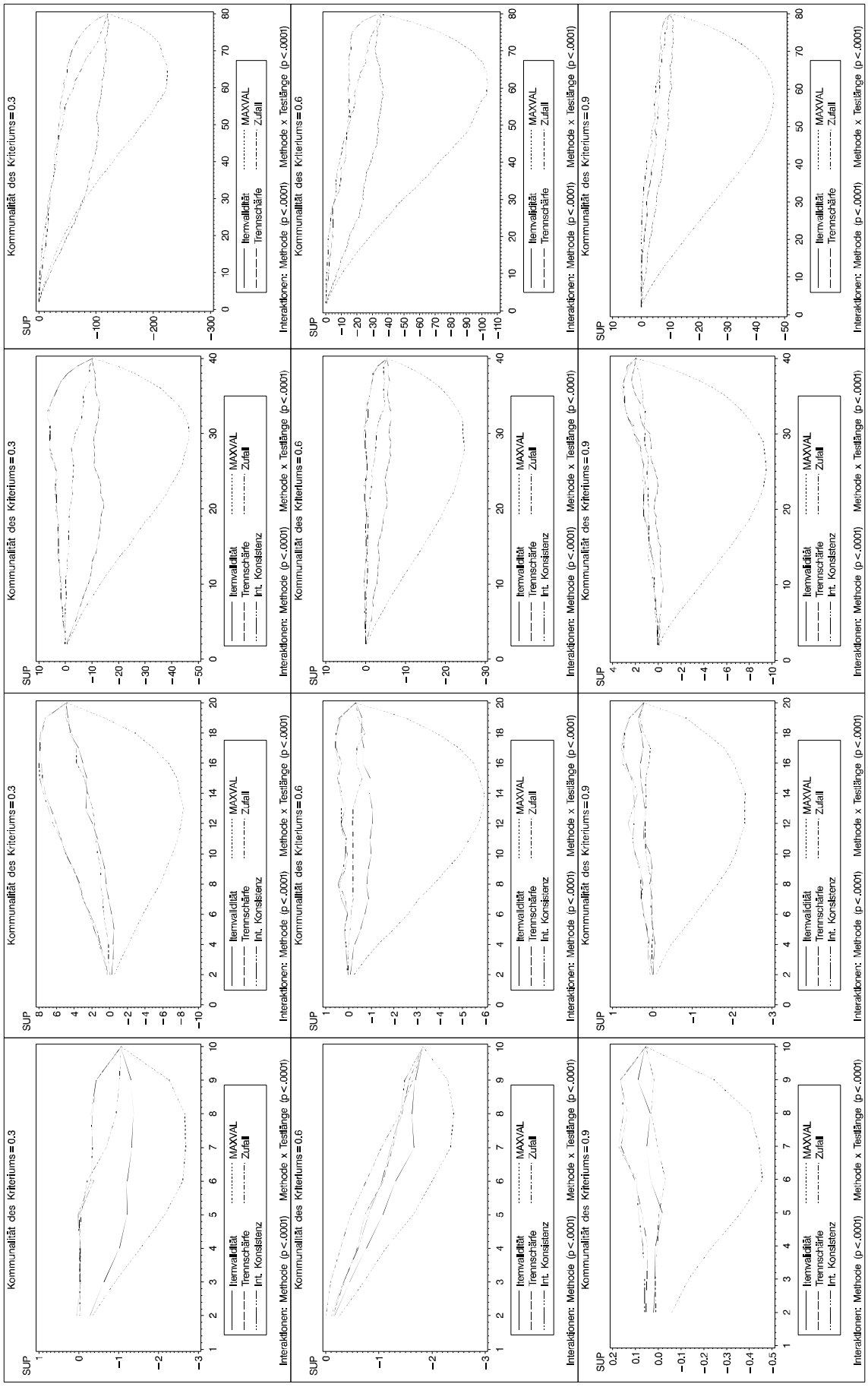


Abbildung 18: Interaktion Kommunalität des Kriteriums mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Kovarianz der Residuen in der Stichprobe (Fisher Z-transformiert) in Studie 2 (getrennte Darstellung je nach Umfang des Itempools).



5.2.3 Studie 3

In der dritten Simulationsstudie wurde untersucht, wie die Selektionsverfahren bei mehrdimensionalen Itempools abschneiden. Dabei wurde in den Studien 3a, 3b und 3c jeweils eine andere Verteilung von valider Varianz gemeinsamer Fehlervarianz und spezifischer Varianz vorgenommen (vgl. Tabelle 3 auf S. 66). Da ansonsten alle anderen Faktoren in allen drei Studien in gleicher Weise variiert wurden, lassen sich die Ergebnisse gemeinsam analysieren, indem man die Verteilung zwischen den einzelnen Varianzkomponenten als Faktor (Stufen: Studie 3a, b oder c) in das Design mit aufnimmt. In Tabelle 5 sind die p-Werte des Haupteffekts der Selektionsmethode sowie deren Interaktion mit den anderen Prädiktoren bei Vorhersage der Validität in der Population dargestellt. Effekte, deren p-Wert unabhängig von Umfang des Itempools größer als ein Promille ist, wurden nicht in Tabelle 5 aufgenommen.

Tabelle 5: P-Werte des Haupteffekts der Selektionsmethode sowie dessen Interaktion mit den anderen Prädiktoren bei Vorhersage der Validität in der Population in Studie 3.

Umfang des Itempools	10	20	40	80
Methode	<.0001	<.0001	<.0001	<.0001
Methode * Studie	<.0001	<.0001	<.0001	<.0001
Methode * Komkrit	<.0001	<.0001	<.0001	<.0001
Methode * Studie * Komkrit	<.0001	<.0001	<.0001	<.0001
Methode * Dimensionen	<.0001	<.0001	<.0001	<.0001
Methode * Studie * Dimensionen	<.0001	<.0001	<.0001	<.0001
Methode * Komkrit * Dimensionen	.0003	<.0001	<.0001	<.0001
Methode * Studie * Komkrit * Dimensionen	.0010	.2452	<.0001	<.0001
Methode * N	<.0001	<.0001	<.0001	<.0001
Methode * Komkrit * N	.6554	.1437	.9252	<.0001
Testlänge * Methode	<.0001	<.0001	<.0001	<.0001
Testlänge * Methode * Studie	<.0001	<.0001	<.0001	<.0001
Testlänge * Methode * Komkrit	<.0001	<.0001	<.0001	<.0001
Testlänge * Methode * Studie * Komkrit	<.0001	.0035	<.0001	<.0001
Testlänge * Methode * Dimensionen	<.0001	<.0001	<.0001	<.0001
Testlänge * Methode * Studie * Dimensionen	<.0001	<.0001	<.0001	<.0001
Testlänge * Methode * Komkrit * Dimensionen	.0578	.0031	<.0001	<.0001
Testlänge * Methode * Studie * Komkrit * Dimensionen	.1280	.2261	.0003	<.0001
Testlänge * Methode * N	<.0001	<.0001	<.0001	<.0001
Testlänge * Methode * Studie * N	.6775	.0992	<.0001	<.0001
Testlänge * Methode * Komkrit * N	.0236	.0068	<.0001	<.0001
Testlänge * Methode * Studie * Komkrit * N	.2536	.0058	<.0001	<.0001
Testlänge * Methode * Dimensionen * N	.2899	.0243	<.0001	<.0001
Testlänge * Methode * Studie * Dimensionen * N	.0122	.0358	<.0001	<.0001
Testlänge * Methode * Komkrit * Dimensionen * N	.0589	.0265	<.0001	<.0001
Testlänge * Methode * Studie * Komkrit * Dimensionen * N	.0004	.0006	<.0001	<.0001

Komkrit = Kommunalität des Kriteriums, N =Stichprobenumfang

Man erkennt, dass die Unterschiede der Selektionsverfahren in Studie 3a, b und c nicht in derselben Größenordnung sind. Der Effekt der Selektionsmethode scheint auch in unterschiedlicher Weise von der Testlänge, der Kommunalität des Kriteriums und der Anzahl der latenten Faktoren abzuhängen. Vor allem bei umfangreichen Itempools gibt es darüber hinaus

eine Reihe von höchstsignifikanten höheren Interaktionen der Prädiktoren unter Beteiligung des Faktors Studie. Daher werden die Ergebnisse der drei Studien im Folgenden getrennt dargestellt. Auch die in den Graphiken mitgeteilten p-Werte beziehen sich jeweils auf die Ergebnisse innerhalb einer Teilstudie.

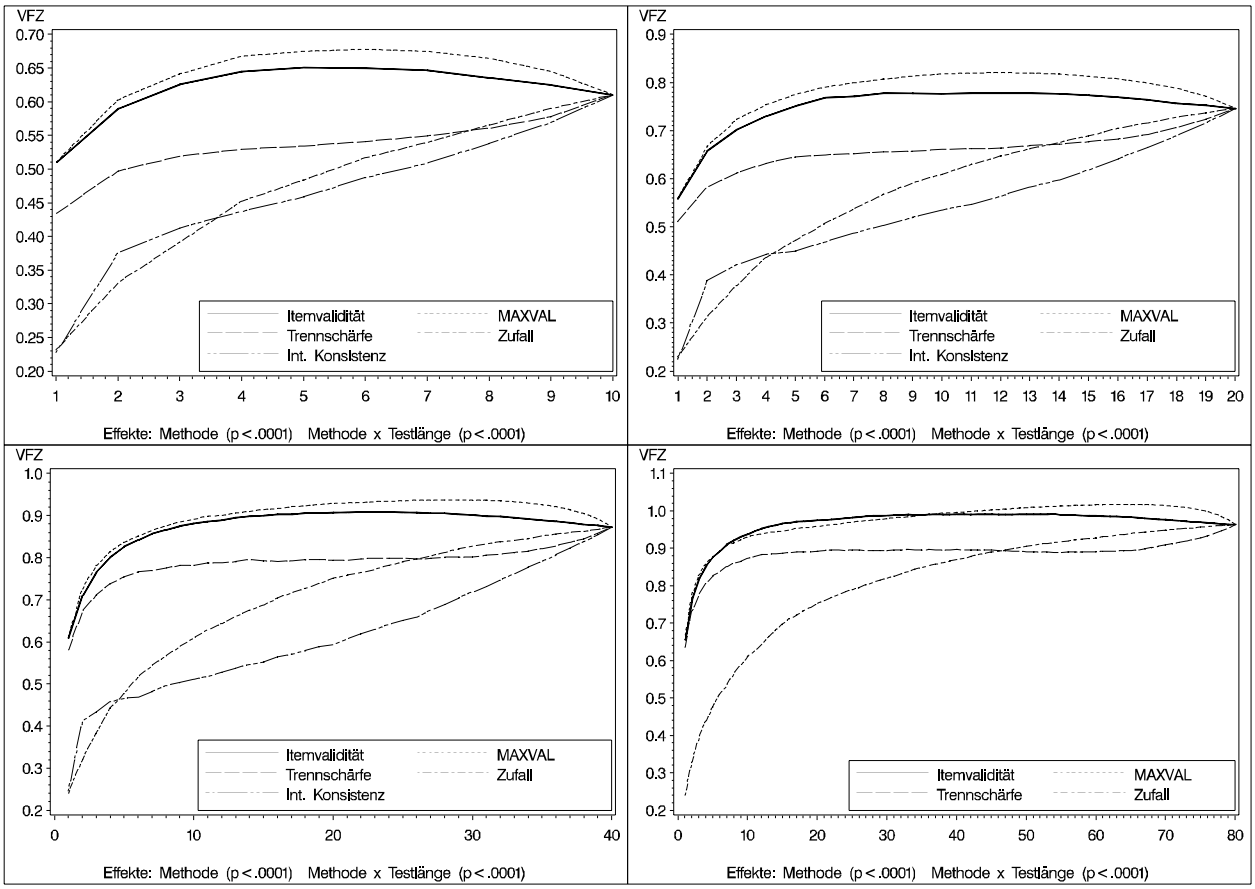
Präsentiert werden vor allem Ergebnisse zur Validität in der Population. Bei signifikanten Unterschieden werden jeweils auch die entsprechenden Ergebnisse zur Validität in der Stichprobe dargestellt. Daneben werden die verschiedenen Selektionsverfahren auch hinsichtlich ihrer Reliabilität verglichen und es wird untersucht, ob bei den resultierenden Skalen Fehlerredundanz oder Fehlersuppression vorliegt (vgl. Kapitel 3.2). Signifikante höhere Interaktionen werden dabei nur dann besprochen, wenn sich Konsequenzen für die Interpretation der niederen Interaktionen erkennen lassen.

5.2.3.a Studie 3a

Validität

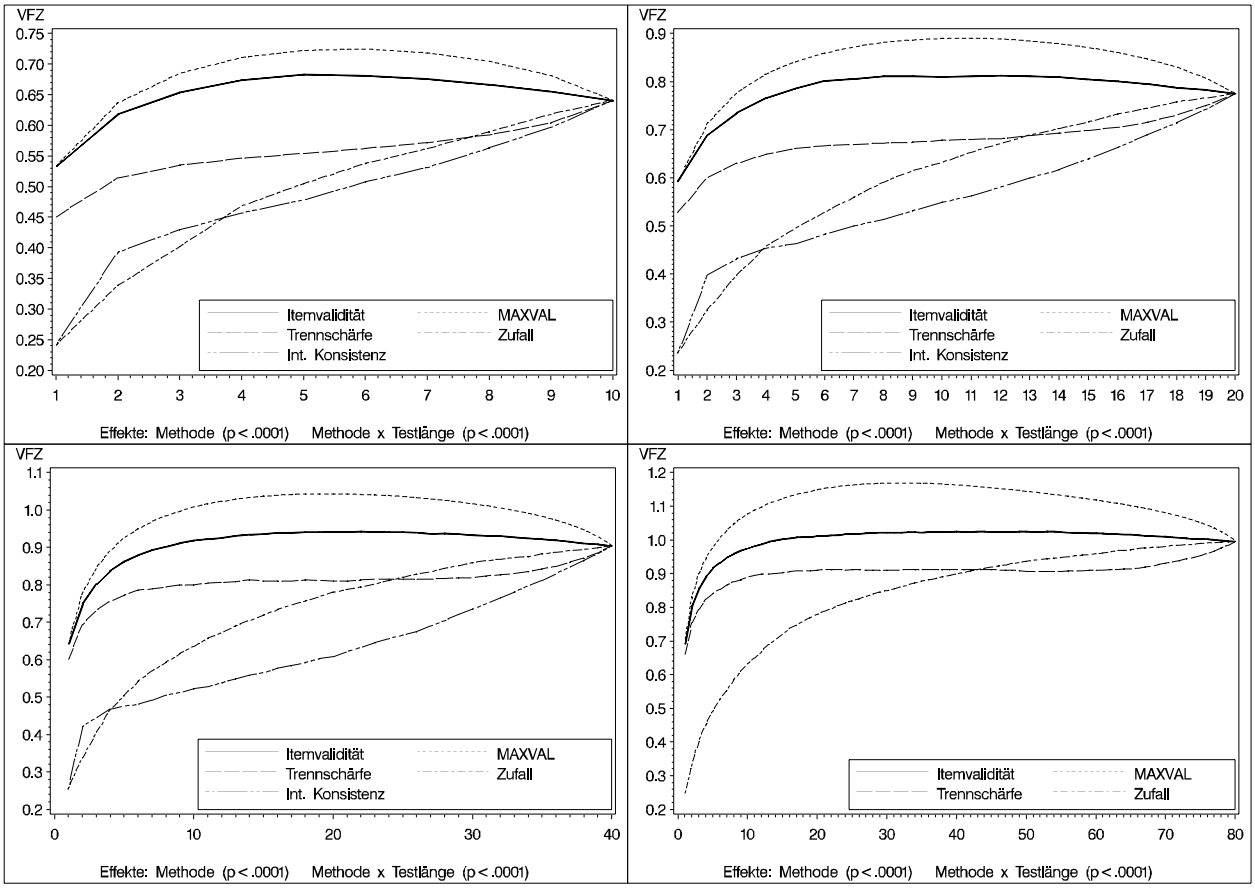
Haupteffekt der Methode und deren Interaktion mit der Testlänge

Abbildung 19: Validität der verschiedenen Selektionsverfahren in Studie 3a (in der Population)



In Studie 3a schnitten die Verfahren, die auf Validitätsdaten basieren (MAXVAL, Itemvalidität) in der Population besser ab als die Verfahren, die auf eine Homogenisierung des Itempools (Trennschärfe, Optimierung von Cronbachs α) abzielen (vgl. Abbildung 19 auf S. 92). Die Optimierung der internen Konsistenz führte sogar zu schlechteren Ergebnissen als die zufällige Auswahl, wenn nicht nur wenige Items in den Test aufgenommen wurden. Auch die Trennschärfe war der zufälligen Auswahl nur bei kurzen Skalen als Selektionsverfahren überlegen. Wenn nur wenige trennscharfe Items in den Test aufgenommen wurden, resultierten durchaus Tests mit hoher Validität. Entfernt man dagegen lediglich die Items mit der geringsten Trennschärfe, so resultieren Tests, die weniger valide sind als bei zufälliger Auswahl, wenn mehr als etwa 55% (bei 80 Items) bis drei Viertel (bei 10 Items) der Items in den Test aufgenommen wurden.

Abbildung 20: Validität der verschiedenen Selektionsverfahren in Studie 3a (in der Stichprobe)



Die Selektion anhand der Itemvalidität scheint nur dann besser zu sein als das MAXVAL-Verfahren, wenn nur wenige Items aus einem umfangreichen Itempool in den Test aufgenommen werden. Ansonsten führt das MAXVAL-Verfahren zu Tests höherer Validität. In der Stichprobe erreicht das MAXVAL-Verfahren jedoch gerade bei der Auswahl von wenigen Items aus einem umfangreichen Itempool die spektakulärsten Ergebnisse (vgl. Abbildung 20). Ansonsten

stimmen die Ergebnisse in der Stichprobe weitgehend mit den Verhältnissen in der Population überein.

Interaktionen der anderen Haupteffekte mit der Selektionsmethode

Die signifikanten Interaktionen der Selektionsmethode mit den anderen Faktoren des Versuchsplans weisen jedoch darauf hin, dass der Vergleich der einzelnen Selektionsverfahren je nach Ausprägung der anderen Faktoren zu unterschiedlichen Ergebnissen führt.

Die Interaktion der Selektionsmethode mit dem Stichprobenumfang ist nur bei umfangreicheren Itempools signifikant. Dies dürfte damit zusammenhängen, dass das MAXVAL-Verfahren bei kleineren Stichprobenumfängen in der Population weniger valide ist und zwar vor allem dann, wenn nur ein kleiner Teil der Items in den Test aufgenommen wird (vgl. Abbildung 21 auf S. 95). In der Stichprobe erreicht das MAXVAL-Verfahren jedoch gerade bei der Selektion von wenigen Items aus umfangreichen Itempools in kleinen Stichproben besonders spektakuläre Validitäten (vgl. Abbildung 22 auf S. 96). Erwartungsgemäß verringern sich jedoch mit zunehmendem Stichprobenumfang auch bei umfangreichen Itempools die Unterschiede zwischen den Ergebnissen in der Population und der Stichprobe.

Die signifikante Interaktion der Selektionsmethode mit der Kommunalität des Kriteriums (vgl. Abbildung 23 auf S. 97) dürfte darauf zurückzuführen sein, dass die Unterschiede zwischen den einzelnen Selektionsverfahren, wie erwartet, mit zunehmender Kommunalität des Kriteriums ansteigen, wie man an der Skalierung der Ordinaten erkennt. Auch die signifikante Dreifachinteraktion von Selektionsmethode, Testlänge und Kommunalität des Kriteriums dürfte damit zusammenhängen, dass die Unterschiede zwischen den Selektionsverfahren bei einer hohen Kommunalität des Kriteriums besonders drastisch sind, wenn nur ein kleiner Teil der Items in den Test aufgenommen wurde. An der Rangfolge der Selektionsmethoden ändert sich dagegen wenig. Lediglich bei einem umfangreichen Itempool scheint das MAXVAL-Verfahren bei einer geringen Kommunalität in der Population zu weniger validen Tests zu führen als die Selektion anhand der Itemvalidität und der Trennschärfe, sofern nur ein kleiner Teil der Items in den Test aufgenommen wird. Unter diesen Bedingungen wird die Validität der vom MAXVAL-Verfahren ausgewählten Tests in der Stichprobe stark überschätzt (vgl. Abbildung 24 auf S. 98).

Abbildung 21: Interaktion des Stichprobenumfangs mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Population (Fisher Z-transformiert) in Studie 3a (getrennte Darstellung je nach Umfang des Itempools).

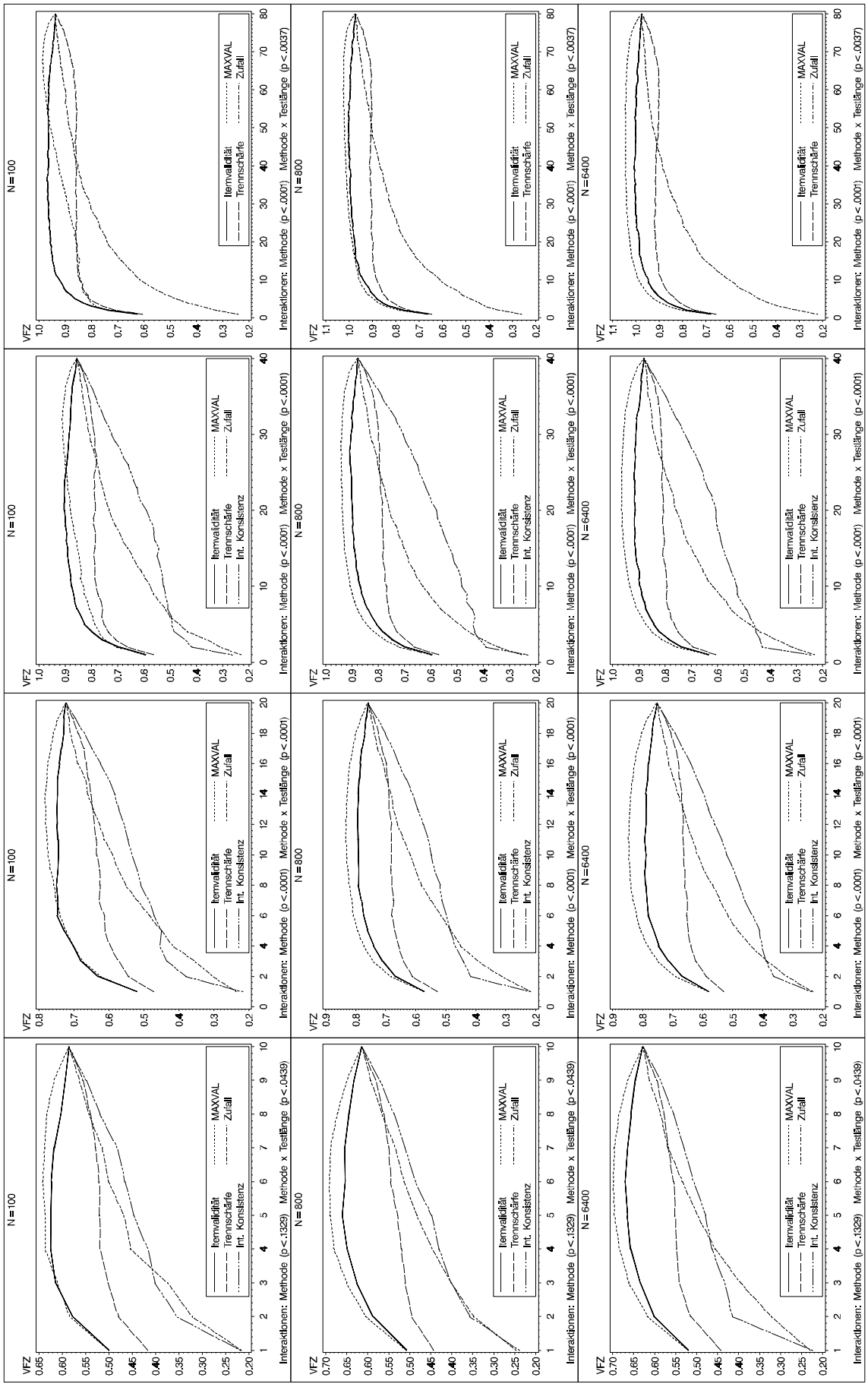


Abbildung 22: Interaktion des Stichprobenumfangs mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Stichprobe (Fisher Z-transformiert) in Studie 3a (getrennte Darstellung je nach Umfang des Itempools).

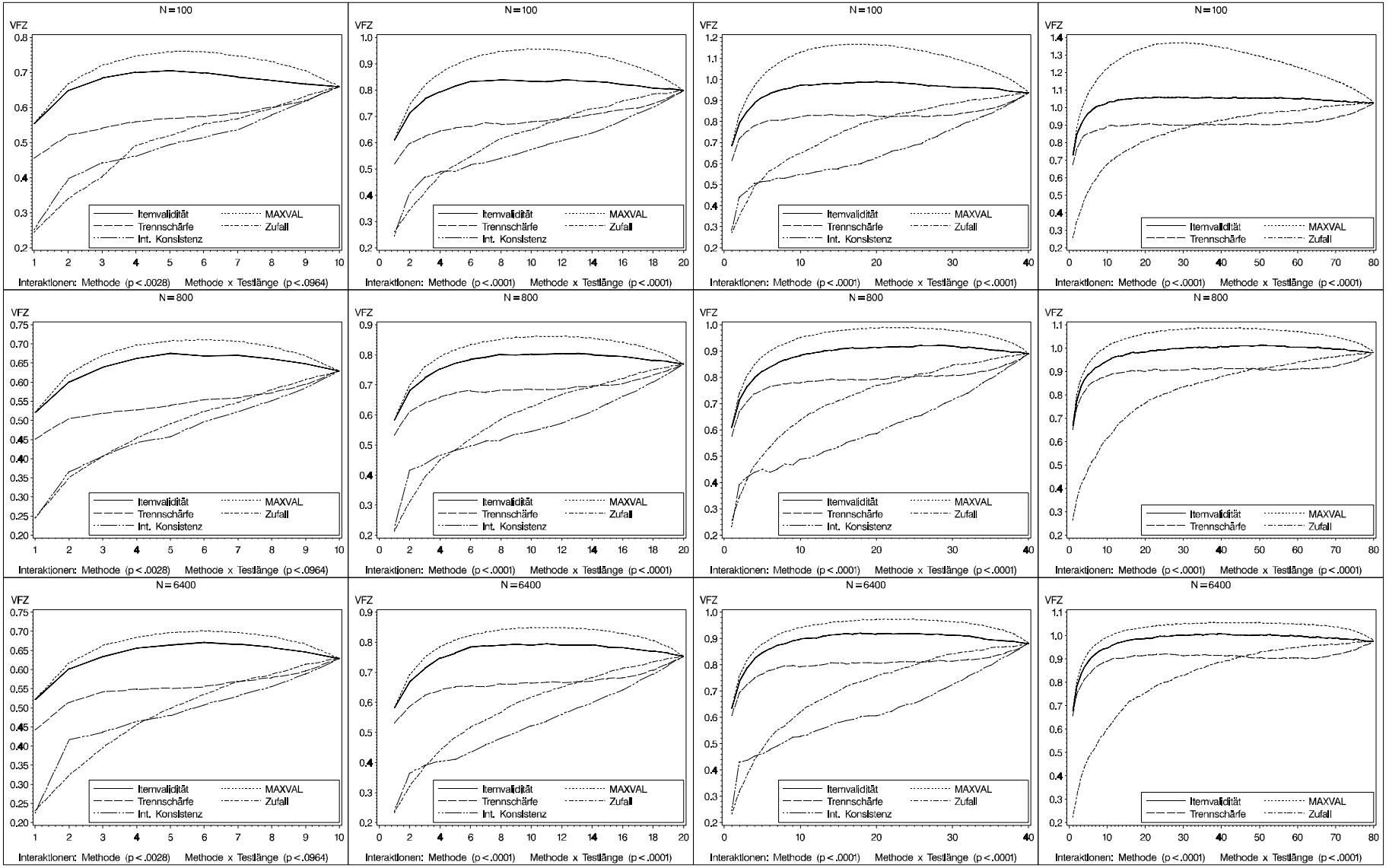


Abbildung 23: Interaktion der Kommunalität des Kriteriums mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Population (Fisher Z-transformiert) in Studie 3a (getrennte Darstellung je nach Umfang des Itempools).

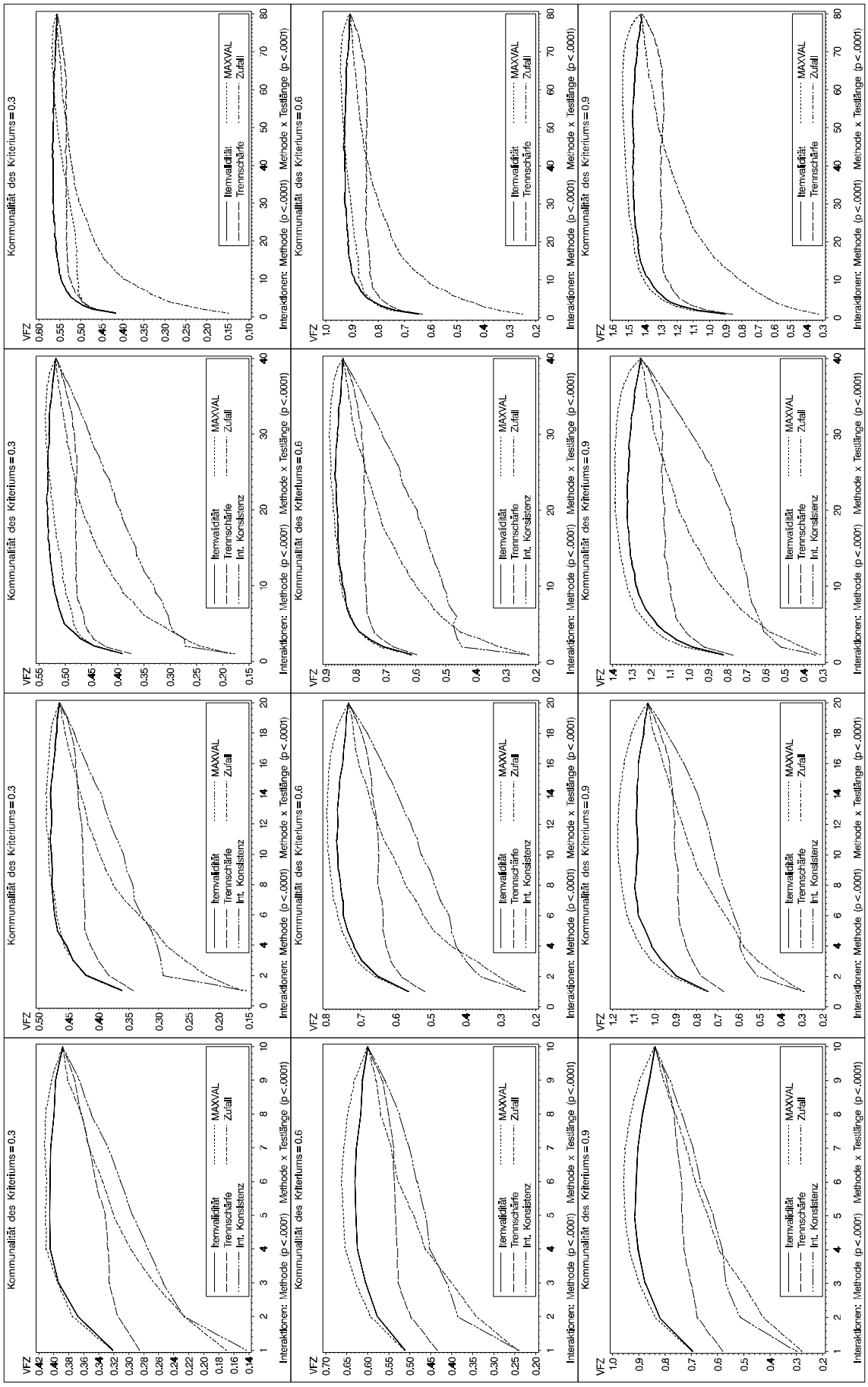


Abbildung 24: Interaktion der Kommunalität des Kriteriums mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Stichprobe (Fisher Z-transformiert) in Studie 3a (getrennte Darstellung je nach Umfang des Itempools).

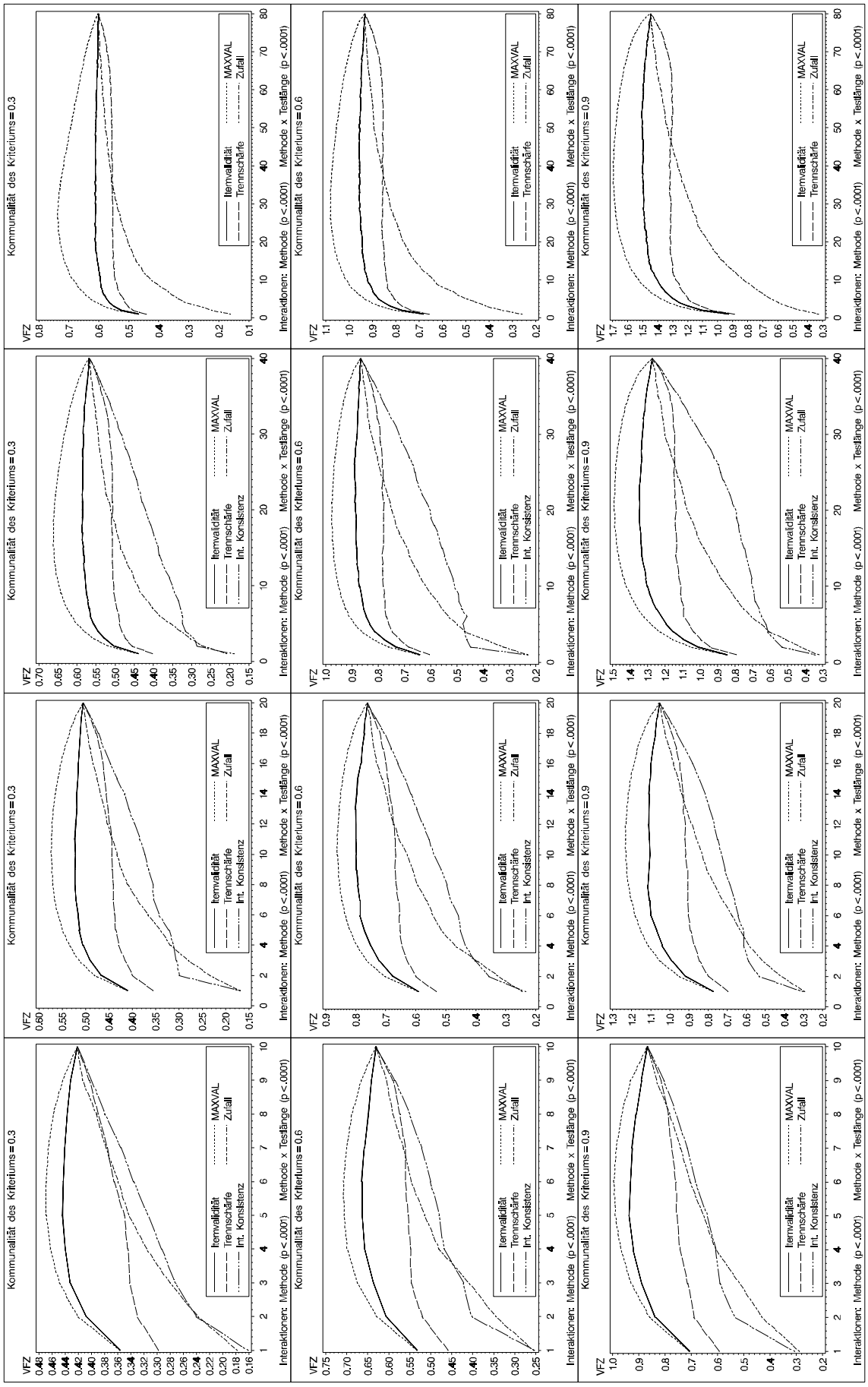


Abbildung 25: Interaktion der Dimensionalität des Itempools mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Population (Fisher Z-transformiert) in Studie 3a (getrennte Darstellung je nach Umfang des Itempools).

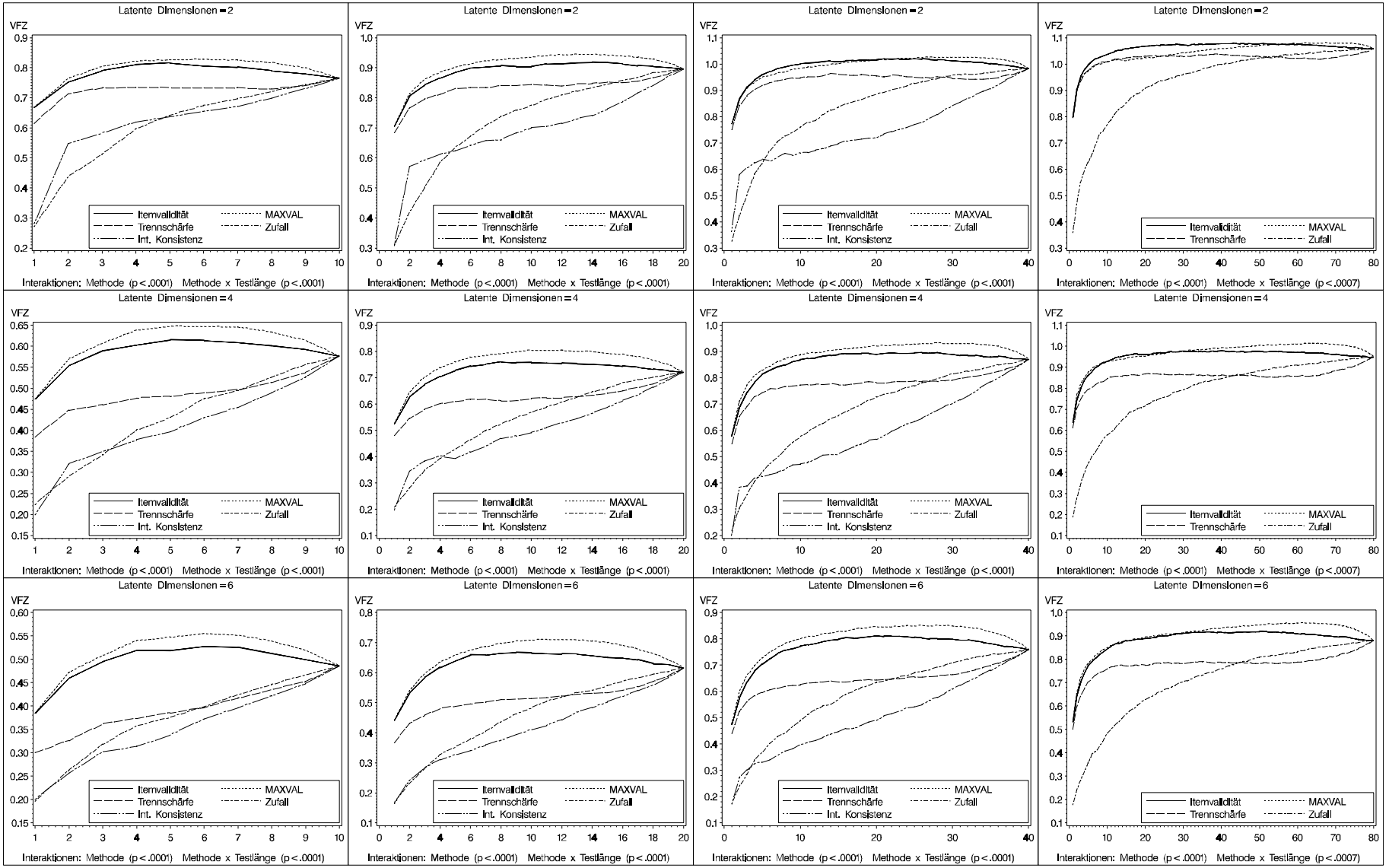


Abbildung 26: Interaktion der Dimensionalität des Itempools mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Stichprobe (Fisher Z-transformiert) in Studie 3a (getrennte Darstellung je nach Umfang des Itempools).

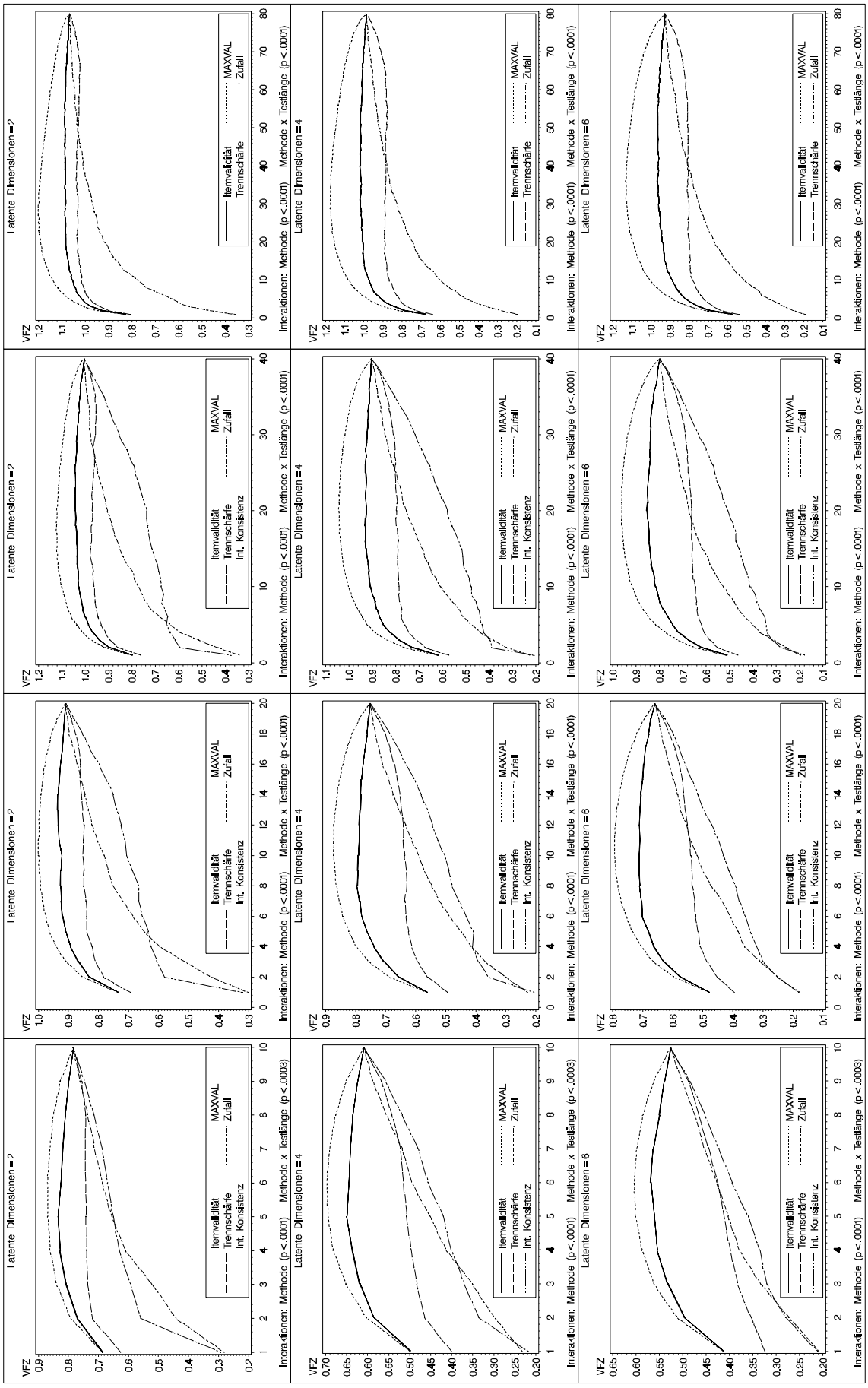


Abbildung 27: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und der Dimensionalität des Itempools bei der Vorhersage der Validität in der Population in Studie 3a (Testlänge 40)

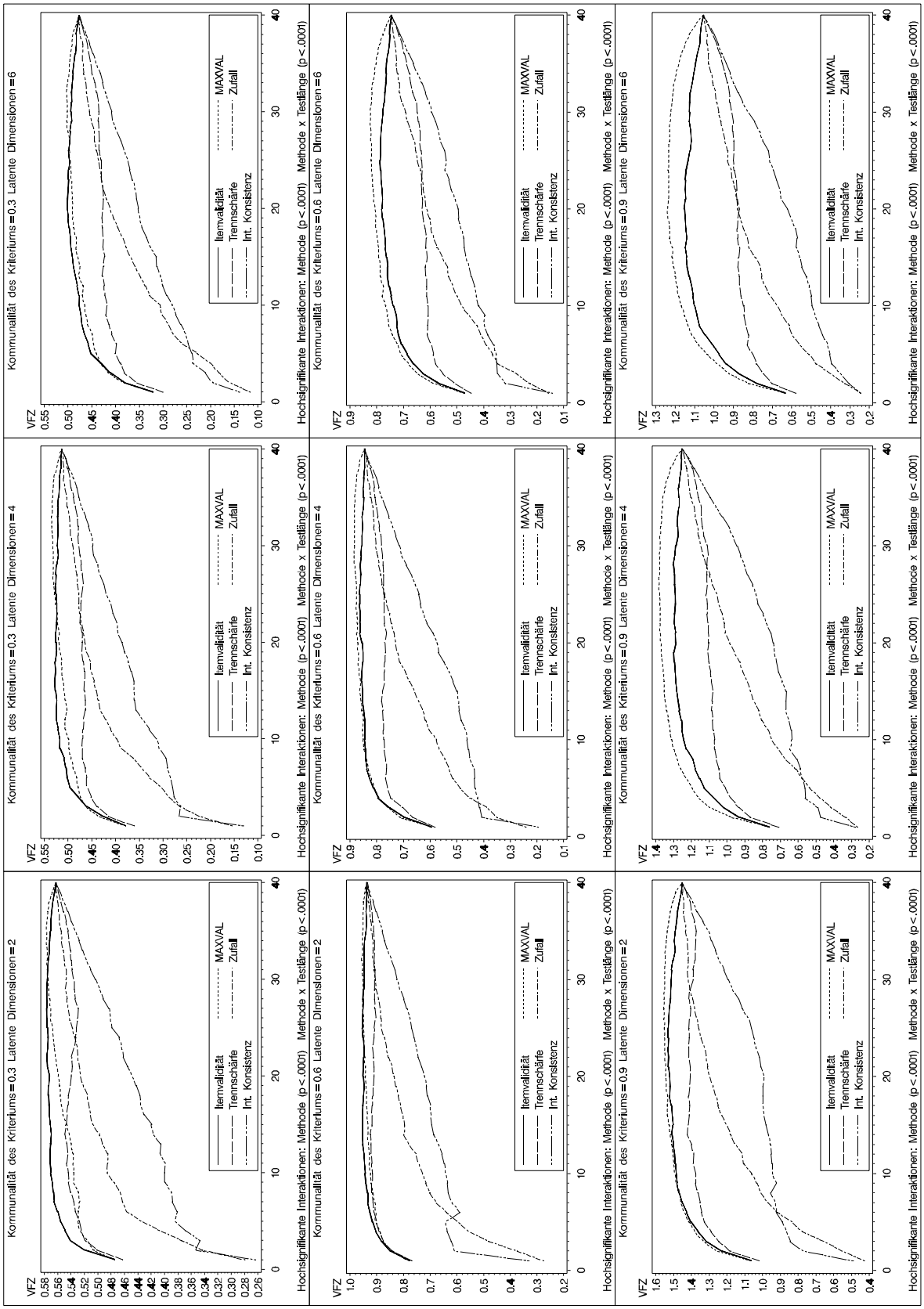
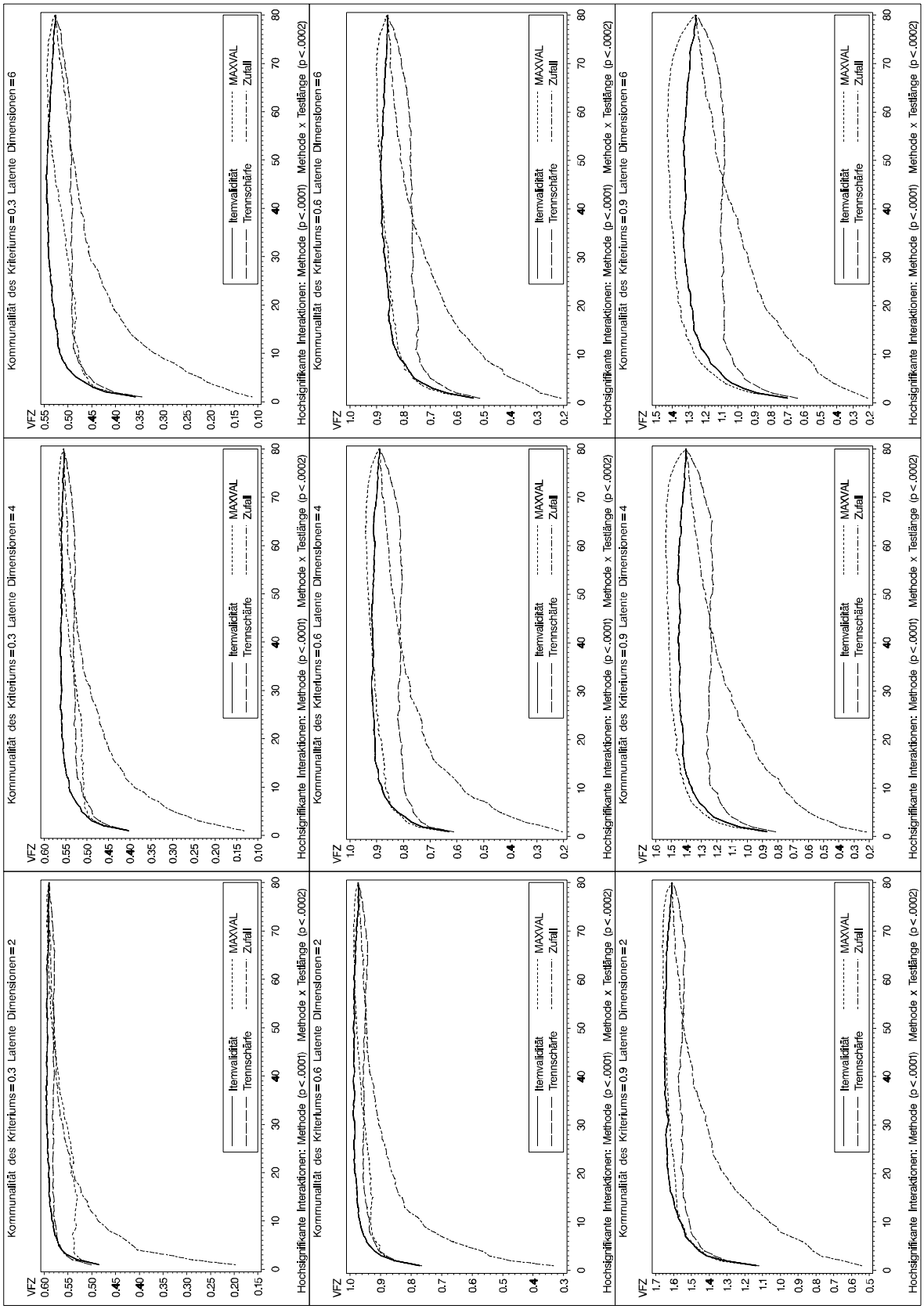


Abbildung 28: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und der Dimensionalität des Itempools bei der Vorhersage der Validität in der Population in Studie 3a (Testlänge 80)

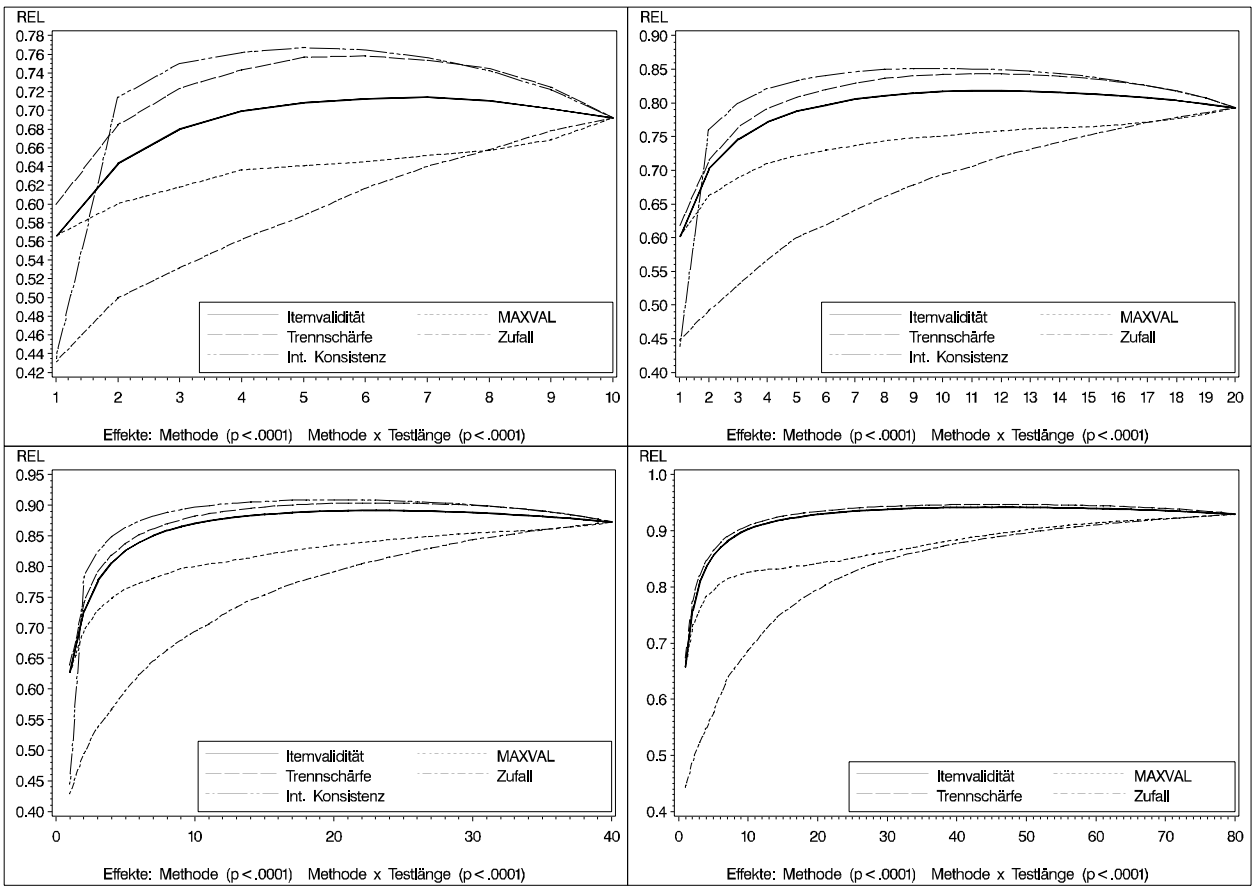


Die signifikante Interaktion der Selektionsmethode mit der Anzahl der Faktoren (vgl. Abbildung 25 auf S. 99 und Abbildung 26 auf S. 100) dürfte darauf zurückzuführen sein, dass die Unterschiede zwischen der Trennschärfe und den beiden Verfahren, die auf Validitätsdaten basieren, in der Population umso größer sind, je mehr latente Faktoren der Itempool hat. Auch in Relation zur Selektion anhand der Itemvalidität schneidet das MAXVAL-Verfahren bei vielen latenten Dimensionen besser ab. In umfangreichen Itempools treten diese Unterschiede bei einer höheren Kommunalität des Kriteriums noch deutlicher hervor, wie man an der signifikanten Interaktion zwischen der Selektionsmethode, der Kommunalität des Kriteriums und der Dimensionalität des Itempools erkennt (vgl. die Skalierung der Ordinaten in Abbildung 27 auf S. 101 und Abbildung 28 auf S. 102). Dann ist das MAXVAL-Verfahren sogar deutlich schlechter als die Selektion anhand der Trennschärfe, wenn dem Itempool wenig latente Dimensionen zugrunde liegen und die Kommunalität des Kriteriums gering ist.

Reliabilität

Haupteffekt der Methode und deren Interaktion mit der Testlänge

Abbildung 29: Reliabilität der verschiedenen Selektionsverfahren in Studie 3a



Die Maximierung der internen Konsistenz anhand der Stichprobendaten scheint die beste Methode zur Sicherung der Reliabilität in der Population zu sein (vgl. Abbildung 29 auf S. 103). Die Selektion anhand der Trennschärfe führt zu Skalen mit vergleichbarer Reliabilität. Bei umfangreichen Itempools ist auch die Selektion anhand der Itemvalidität nicht wesentlich schlechter. Das MAXVAL-Verfahren führt zu Skalen deren Reliabilität relativ gering ist, wobei die Unterschiede zu den anderen Verfahren etwas geringer sind, wenn nur ein *sehr* kleiner oder ein *sehr* großer Teil der Items in den Test aufgenommen wurde. Wenn nur wenige Items in den Test aufgenommen wurden, resultieren beim MAXVAL-Verfahren Skalen höherer Reliabilität als bei zufälliger Auswahl. Wenn man von der zufälligen Auswahl absieht, dann leisten diejenigen Verfahren, die für die Sicherung der Validität am wenigsten geeignet erscheinen, bei der Sicherung der Reliabilität die besten Dienste.

Interaktionen der anderen Prädiktoren mit der Selektionsmethode

Die signifikante Interaktion der Selektionsmethode mit der Anzahl der latenten Faktoren sowie die Dreifachinteraktion unter Einschluss der Testlänge scheint darauf zurückzuführen zu sein, dass die Unterschiede zwischen den Selektionsverfahren mit zunehmender Anzahl der latenten Faktoren deutlicher hervortreten. Dieser Effekt ist bei kleineren Itempools sowie bei mittleren Testlängen ausgeprägter (siehe Abbildung 32 auf S. 107).

Außer bei sehr kleinen Itempools werden auch die Interaktionen der Selektionsmethode mit dem Stichprobenumfang (Abbildung 30 auf S. 105) sowie mit der Kommunalität des Kriterium (Abbildung 31 auf S. 106) signifikant. Auch die Dreifachinteraktion dieser Prädiktoren ist dann jeweils signifikant (siehe Abbildung 33, S. 108 bis Abbildung 35, S. 110). Dies scheint darauf zurückzuführen zu sein, dass das MAXVAL-Verfahren bei einer geringen Kommunalität des Kriteriums nur dann wesentlich schlechter abschneidet als die anderen Selektionsverfahren, wenn die Stichprobe klein ist. Bei umfangreichen Stichproben hängen die Unterschiede zwischen den Selektionsmethoden dagegen kaum von der Kommunalität des Kriteriums ab (siehe Abbildung 33, S. 108 bis Abbildung 35, S. 110). Bei umfangreichen Itempools und großen Personenstichproben ist das MAXVAL-Verfahren fast genauso reliabel wie die anderen Methoden (siehe Abbildung 30 auf S. 105 sowie Abbildung 33, S. 108 bis Abbildung 35, S. 110).

Abbildung 30: Interaktion des Stichprobenumfangs mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Reliabilität in Studie 3a (getrennte Darstellung je nach Umfang des Itempools).

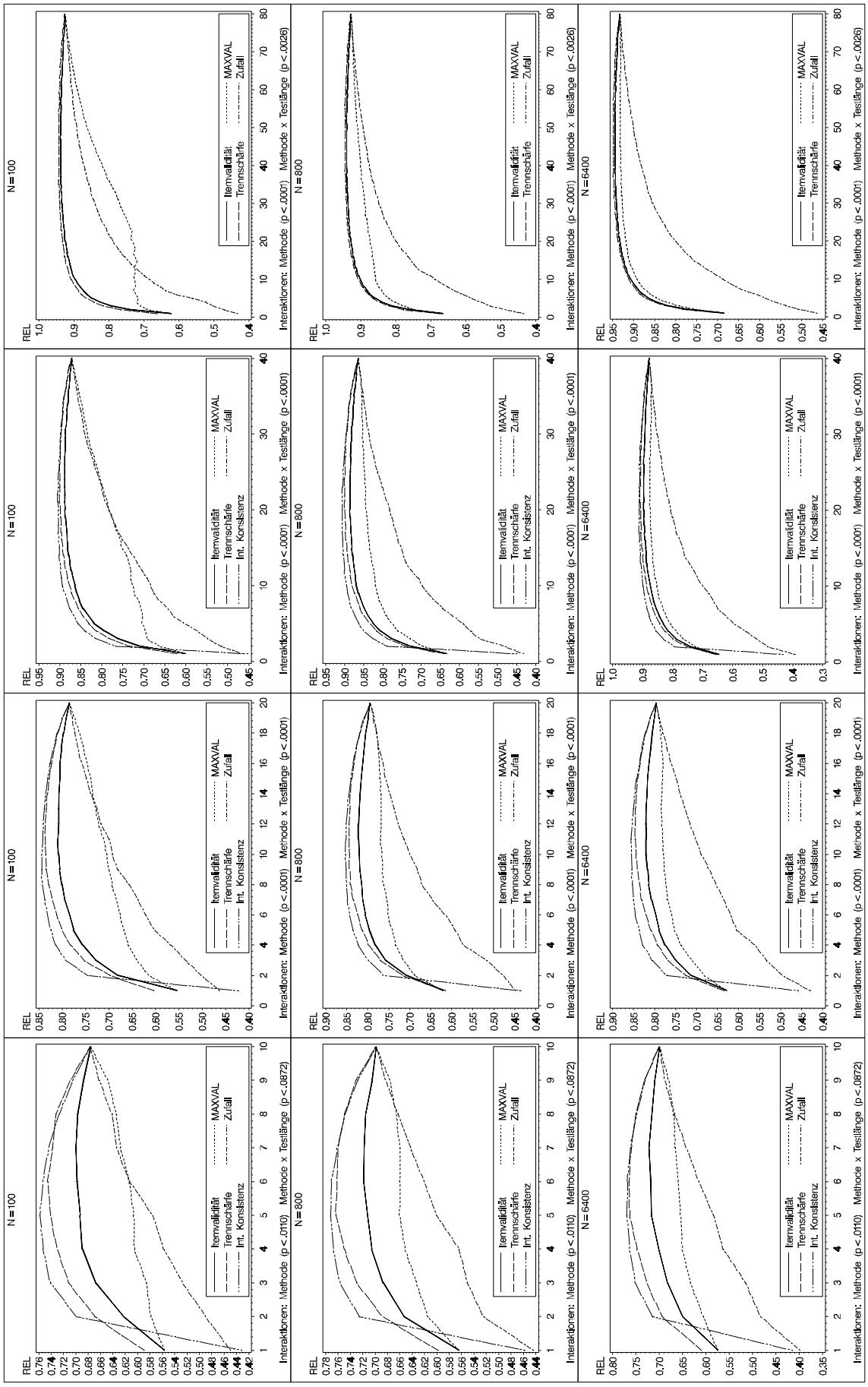


Abbildung 31: Interaktion der Kommunalität des Kriteriums mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Reliabilität in Studie 3a (getrennte Darstellung je nach Umfang des Itempools).

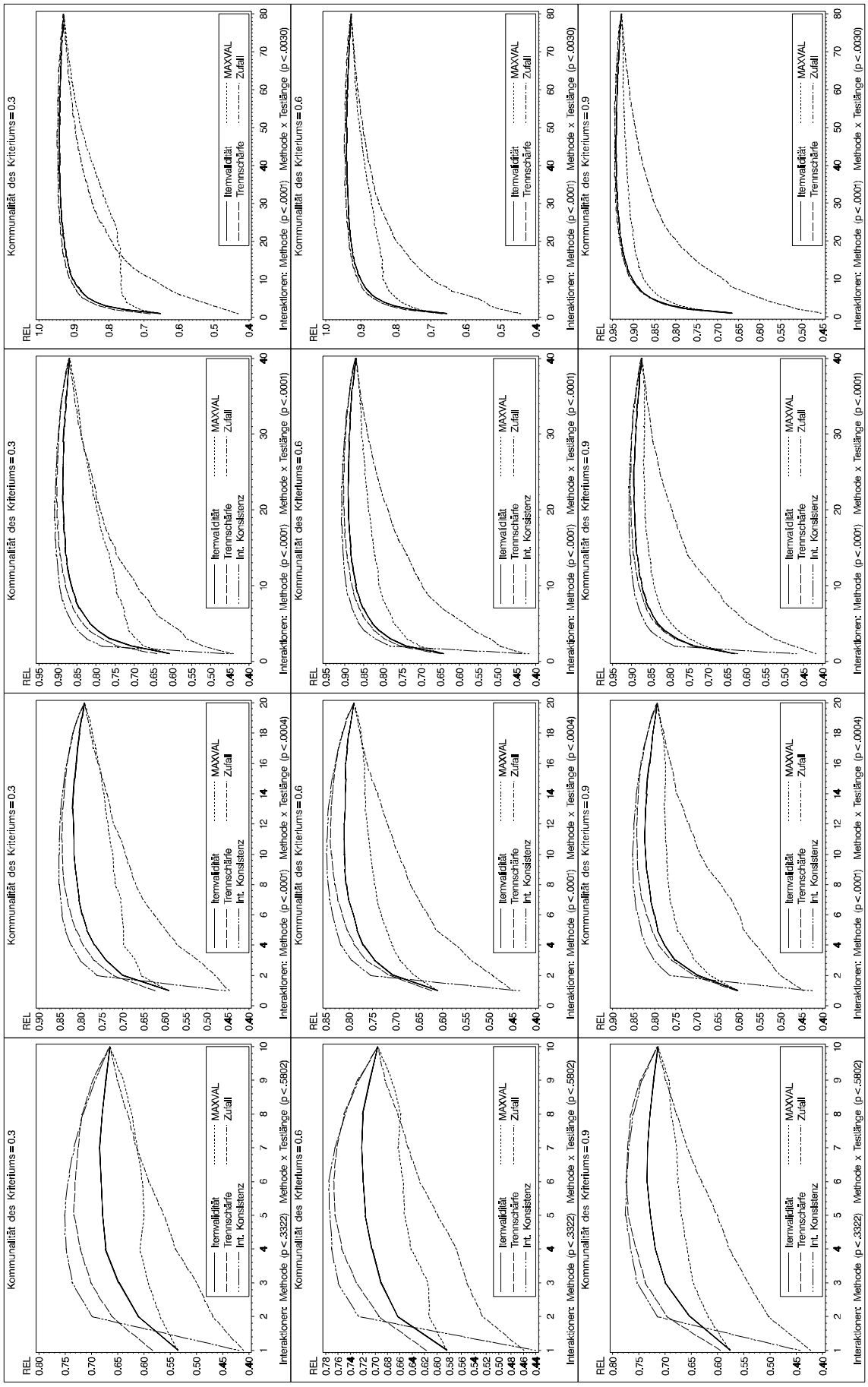


Abbildung 32: Interaktion der Dimensionalität des Itempools mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Reliabilität in Studie 3a (getrennte Darstellung je nach Umfang des Itempools).

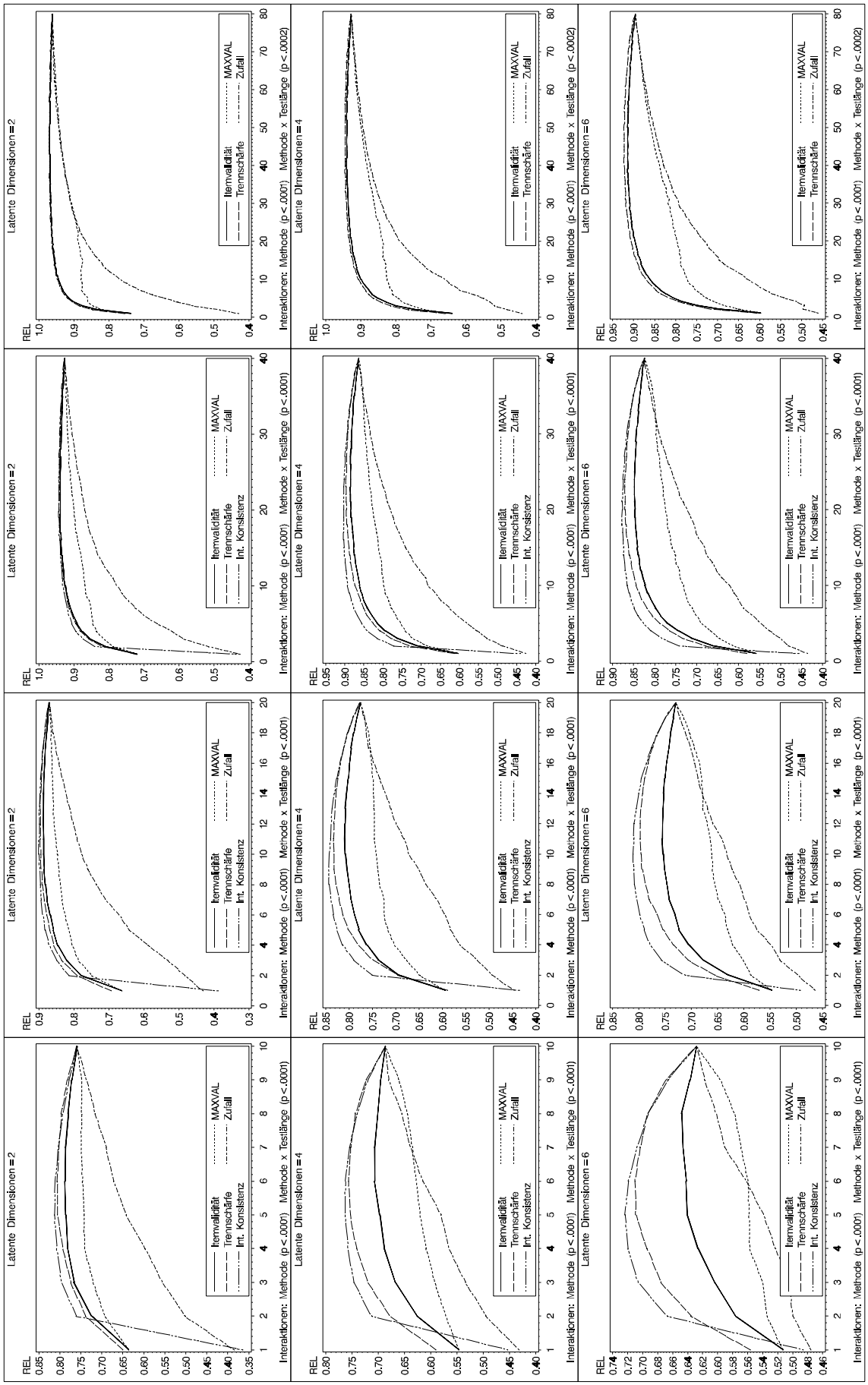


Abbildung 33: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Reliabilität in Studie 3a (Testlänge 20)

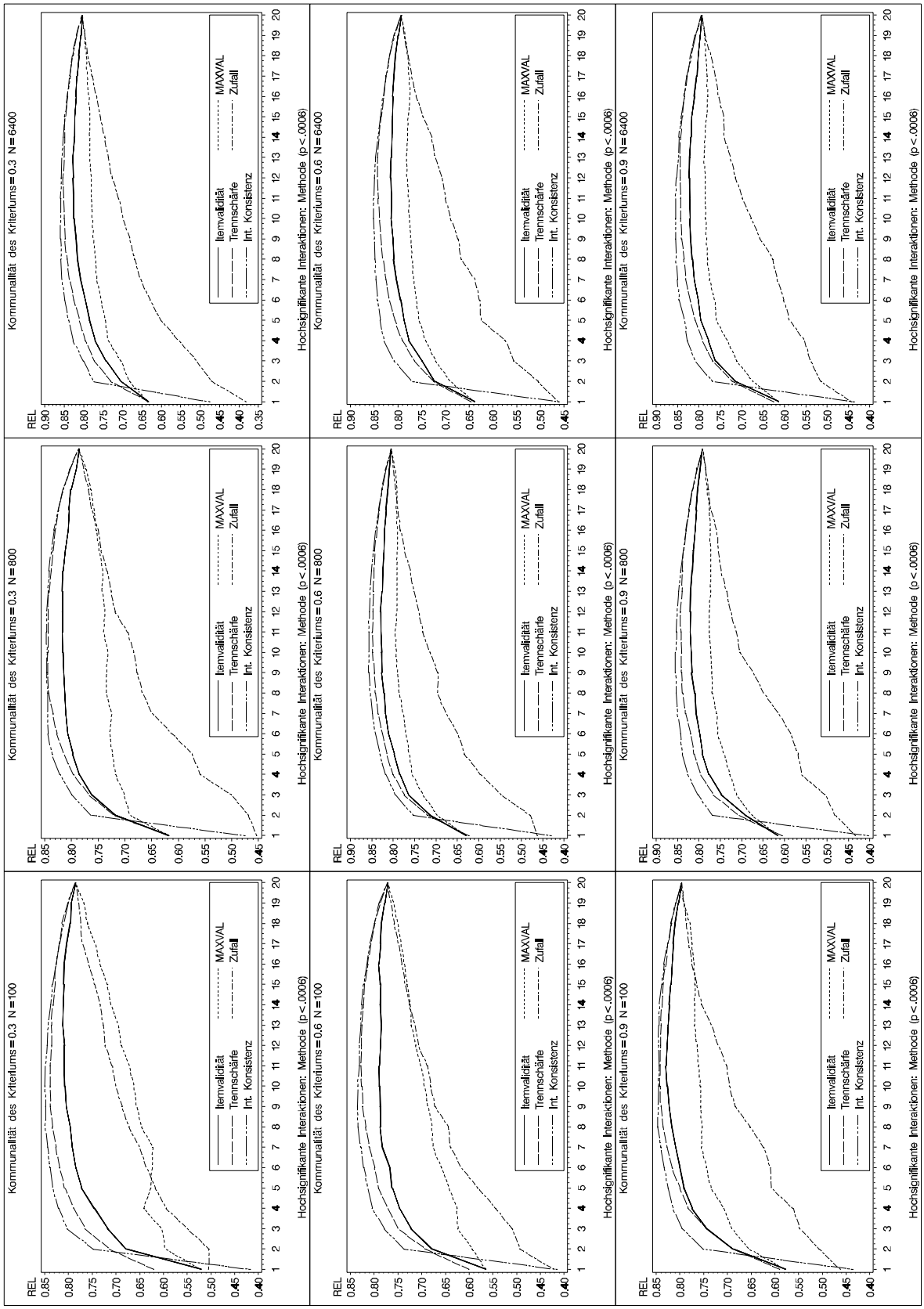


Abbildung 34: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Reliabilität in Studie 3a (Testlänge 40)

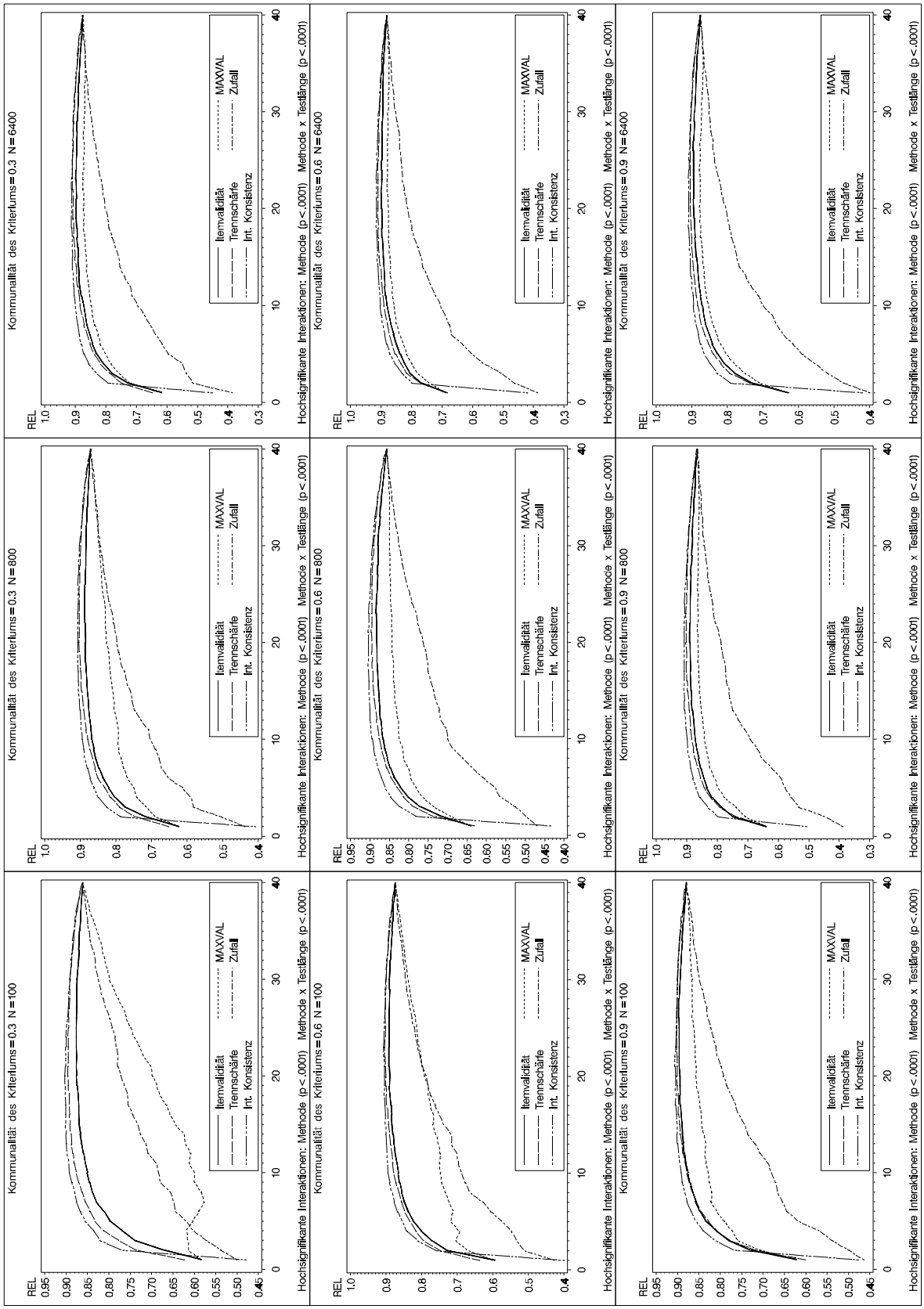
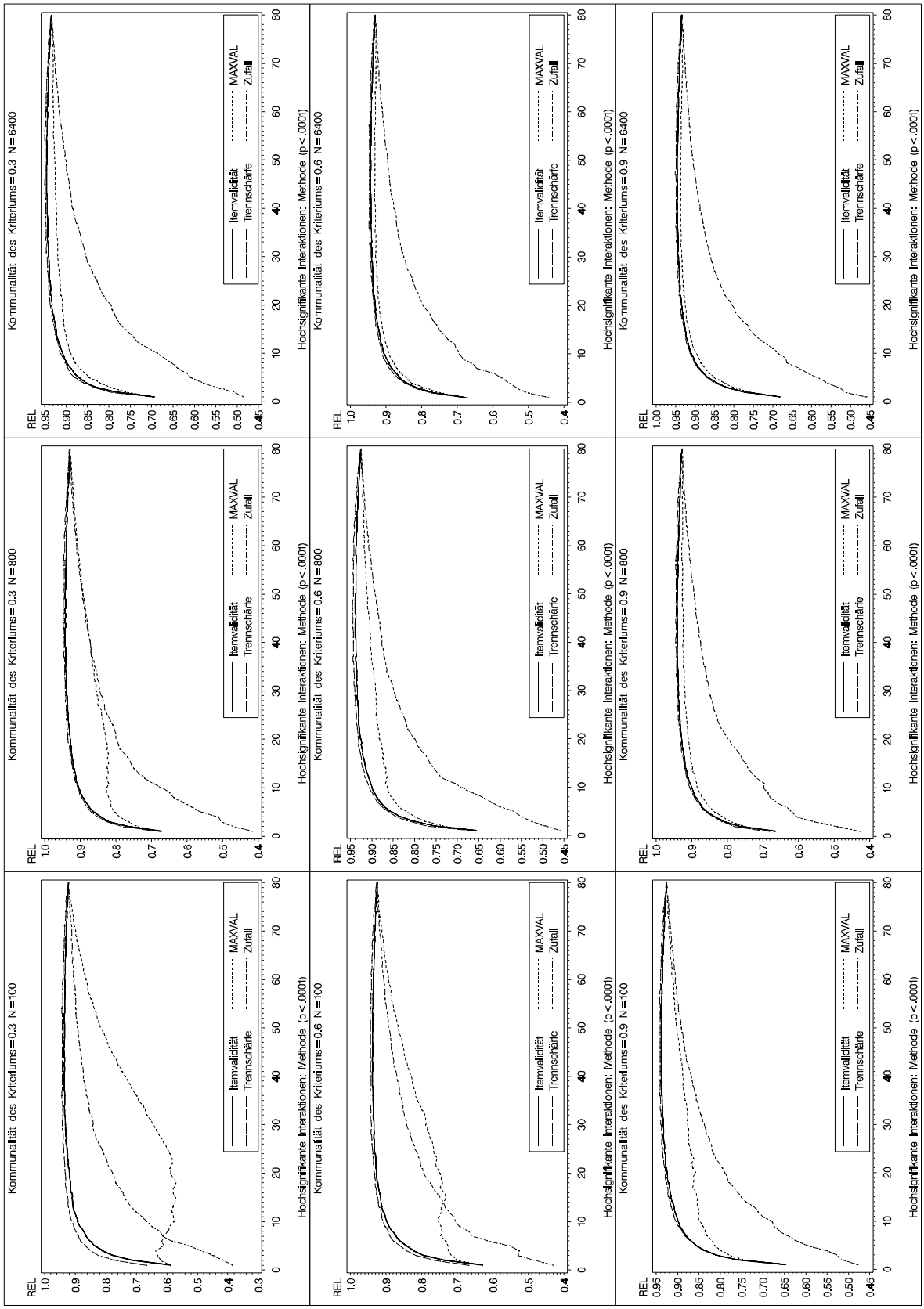


Abbildung 35: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Reliabilität in Studie 3a (Testlänge 80)

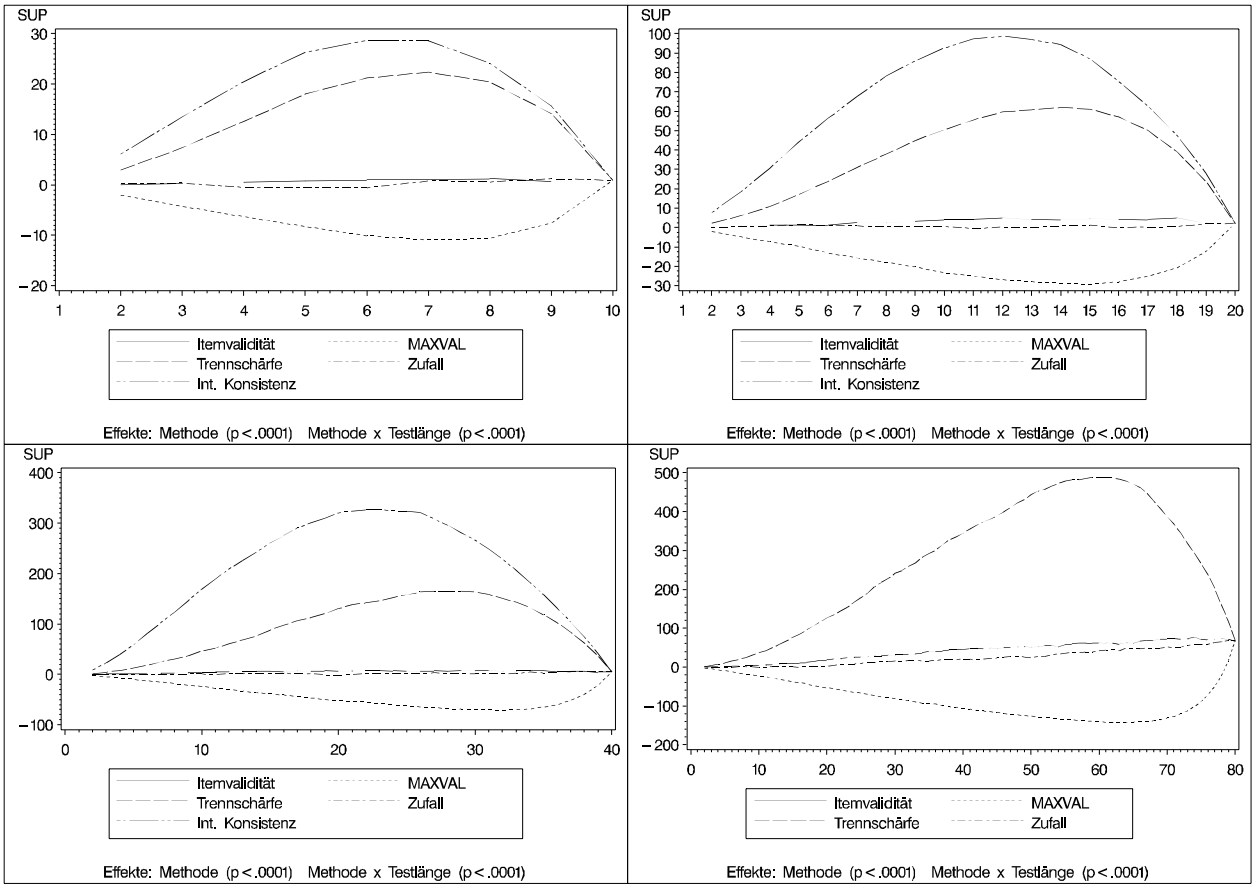


Suppression

Haupteffekt

In Kapitel 3.2 wurde ein neues Kriterium für Suppression in der Testkonstruktion eingeführt. Demnach liegt Fehlersuppression genau dann vor, wenn (die Summe der) Kovarianzen der Residualanteile der Items bei Herauspartialisierung der wahren Werte des Kriteriums negativ sind (vgl. [3.2-5] auf S. 44 und [3.2-6] auf S. 44). In Abbildung 36 ist jeweils die Summe der Kovarianzen der Residuen für die von den verschiedenen Selektionsmethoden ausgewählten Tests dargestellt. Bei negativen Werten liegt Fehlersuppression vor, bei positiven Werten Fehlerredundanz. Im Gegensatz zu den Abbildungen in Kapitel 5.2.2 wurde die Kovarianzen der Residuen nicht anhand der Daten in Stichprobe, sondern anhand der Populationskovarianzmatrix berechnet.

Abbildung 36: Kovarianz der Residuen bei Herauspartialisierung der wahren Kriteriumswerte



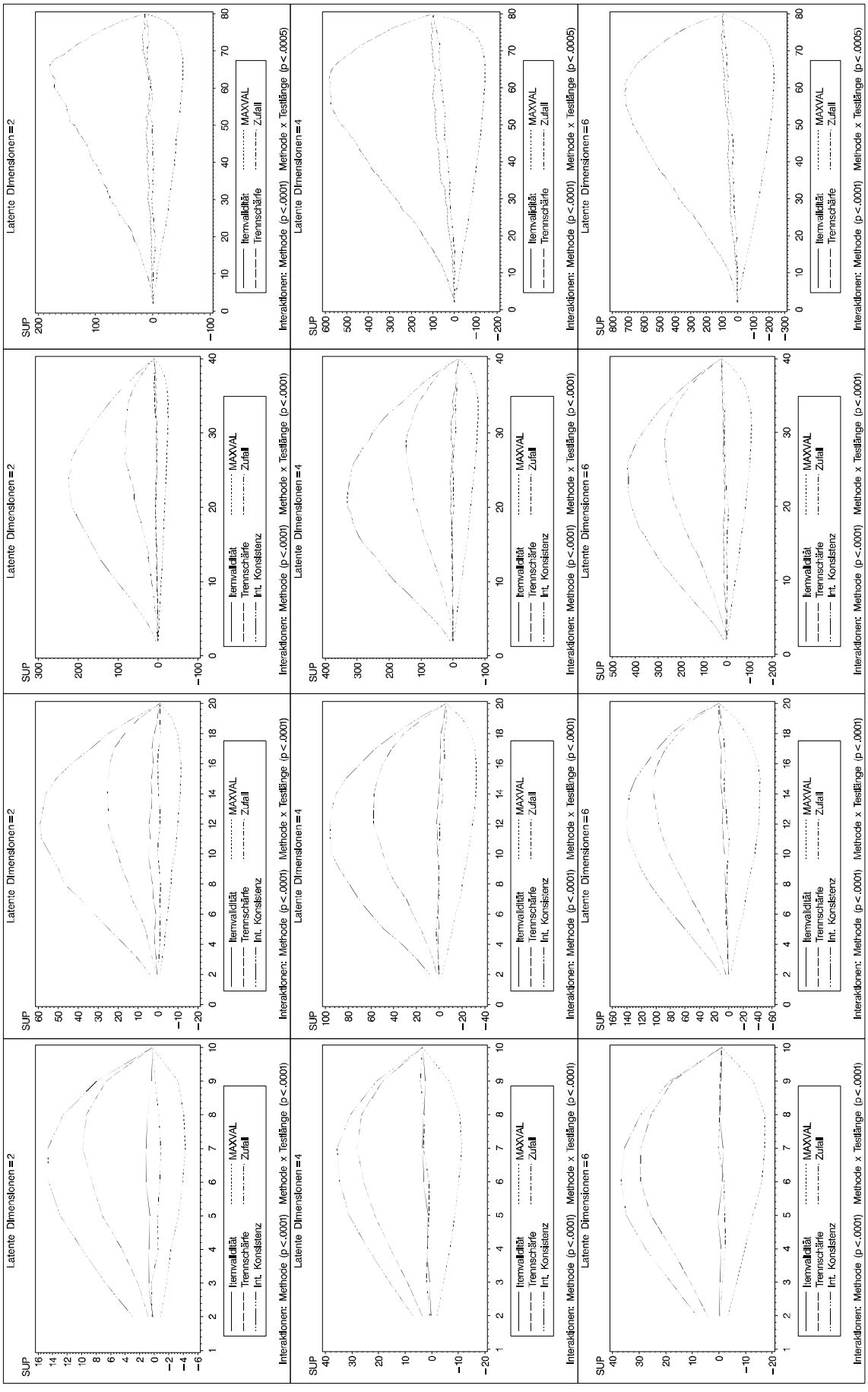
Es zeigt sich, dass bei Verfahren, die eine Maximierung der internen Konsistenz zum Ziel haben, in der Regel Fehlerredundanz vorliegt, während bei einer Optimierung der Validität mit dem MAXVAL-Verfahren Fehlersuppression zu beobachten ist. Die Optimierung von Cronbachs α führt zu deutlich stärkerer Fehlerredundanz als die Selektion anhand der Trennschärfe. Die

Fehlerredundanz ist bei der Optimierung von Cronbachs α am größten, wenn etwa die Hälfte der Items in den Test aufgenommen wurde. Beim MAXVAL-Verfahren sowie der Selektion anhand der Trennschärfe wird das Maximum der Fehlersuppression bzw. Fehlerredundanz erst erreicht, wenn ein größerer Anteil des Itempools in den Test aufgenommen wurde. Je umfangreicher der Itempool ist, desto größer ist der Anteil der Items, die in den Test aufgenommen werden müssen, damit der jeweilige Extremwert erreicht wird.

Interaktionen

Die Unterschiede der verschiedenen Selektionsverfahren hinsichtlich des hier vorgestellten Suppressionskriteriums hängen weder vom Stichprobenumfang noch von der Kommunalität des Kriteriums ab. Es zeigt sich jedoch eine signifikante Interaktion der Selektionsmethode mit der Dimensionalität des Itempools. Je mehr latente Dimensionen dem Itempool zugrunde liegen, desto stärker unterscheiden (=Differenz) sich die Selektionsverfahren hinsichtlich des Suppressionskriteriums, wie man an der Skalierung der Ordinaten der Graphiken in Abbildung 37 (S. 113) erkennt. Der Quotient aus der Fehlerredundanz bei Selektion anhand der Trennschärfe und der Optimierung von Cronbachs α wird jedoch mit zunehmender Anzahl der Faktoren kleiner.

Abbildung 37: Interaktionen der Selektionsmethode mit der Anzahl der Dimensionen des Itempools bei der Vorhersage der Kovarianz der Residuen bei Herauspartialisierung der wahren Kriteriumswerte in Studie 3a (getrennt mach Umfang des Itempools)

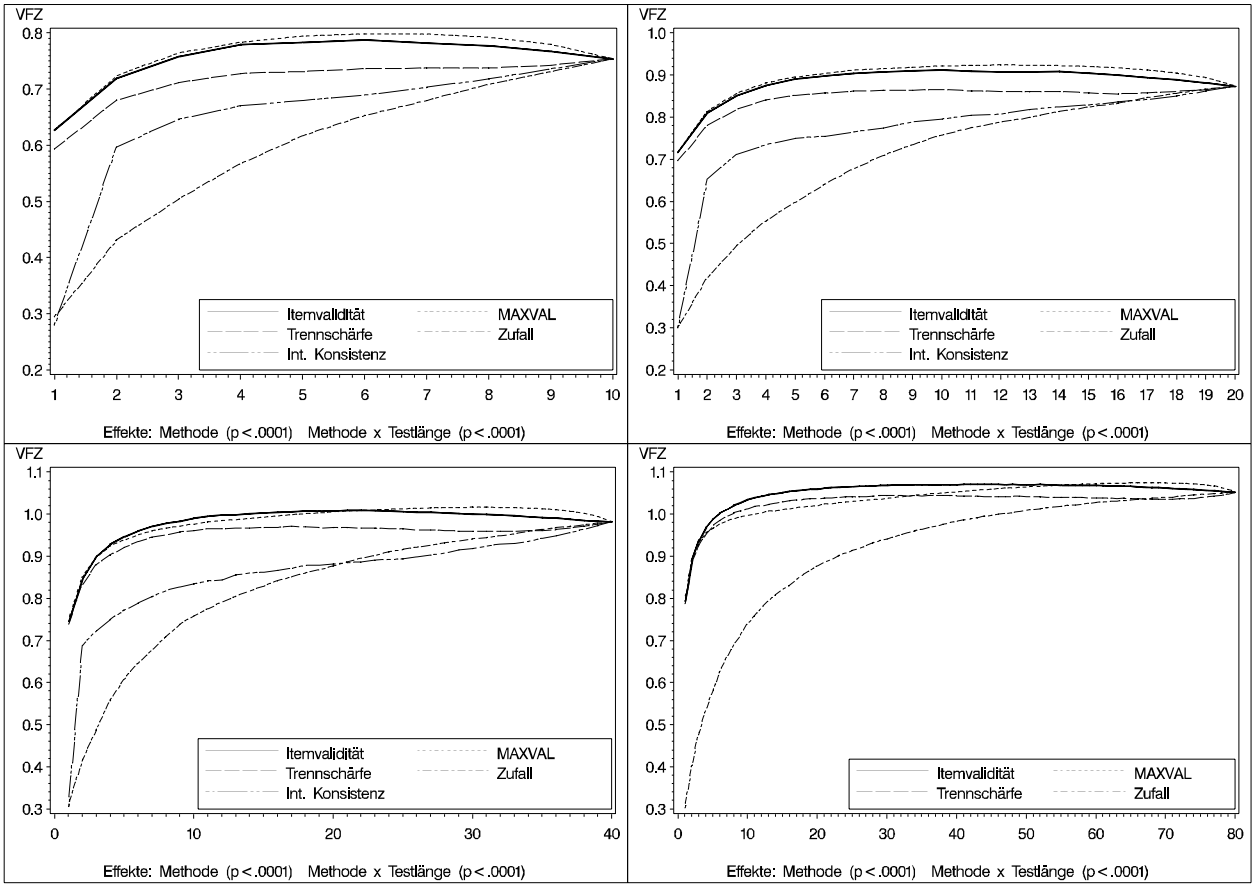


5.2.3.b Studie 3b

Validität

Haupteffekt der Methode und deren Interaktion mit der Testlänge

Abbildung 38: Validität der verschiedenen Selektionsverfahren in Studie 3b (in der Population)

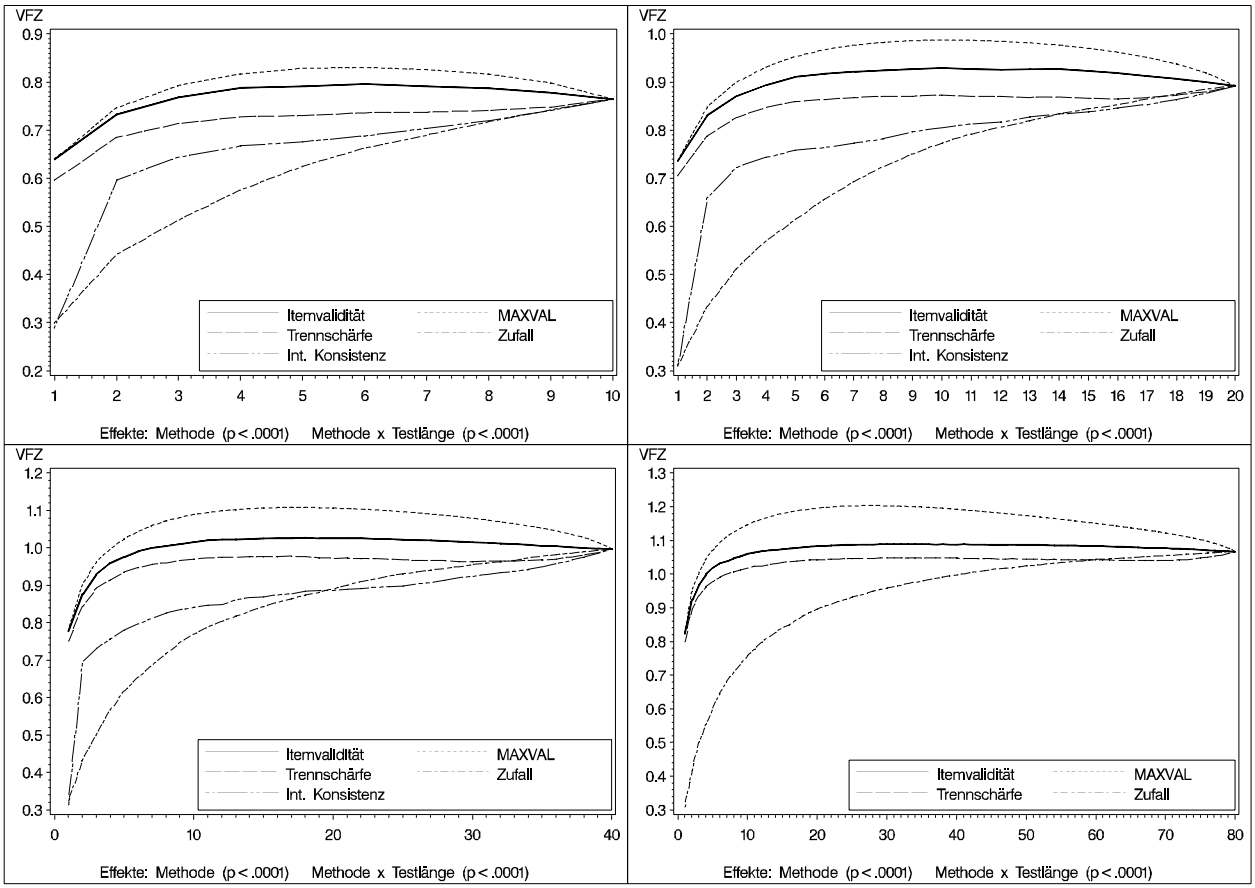


In Studie 3b war das MAXVAL-Verfahren bei umfangreichen Itempools schlechter als die Selektion anhand der Itemvalidität und der Trennschärfe, sofern nur ein kleiner Teil der Items in den Test aufgenommen wurde (vgl. Abbildung 38). Wenn nur wenige Items aus dem Test entfernt werden, ist das MAXVAL-Verfahren jedoch nach wie vor mindestens so gut wie Selektion anhand der Itemvalidität, die wiederum grundsätzlich besser abschneidet als die Selektion anhand der Trennschärfe. Die Unterschiede zwischen der Validität bei Selektion anhand der Trennschärfe und der Validität bei Selektion anhand der Itemvalidität sind jedoch deutlich geringer als in Studie 3a. Die Selektion anhand der Trennschärfe führt in Studie 3b selbst bei umfangreichen Itempools, aus denen ein geringer Teil der Items entfernt wird, nur zu unwesentlich schlechteren Ergebnissen als die zufällige Auswahl, während die Trennschärfe bei kürzeren Tests deutlich besser abschneidet. Außer bei der Entfernung von wenigen Items aus umfangreichen Itempools scheint auch die Optimierung der internen Konsistenz nicht zu Skalen

geringerer Validität zu führen als die zufällige Auswahl. Auch in Studie 3b ist diese Methode jedoch grundsätzlich schlechter als die übrigen Selektionsmethoden.

In der Stichprobe schneidet das MAXVAL-Verfahren so wie in Studie 3a gerade unter den Bedingungen besonders gut ab, bei denen es in der Population besonders schlechte Ergebnisse liefert, nämlich bei der Selektion von wenigen Items aus umfangreichen Itempools (vgl. Abbildung 39). Ansonsten lassen sich auch in Studie 3b nur geringe Unterschiede zwischen den Resultaten der Selektionsverfahren in der Stichprobe und der Population erkennen.

Abbildung 39: Validität der verschiedenen Selektionsverfahren in Studie 3b (in der Stichprobe)



Interaktionen der anderen Haupteffekte mit der Selektionsmethode

Die signifikante Interaktion der Selektionsmethode mit der Kommunalität des Kriteriums sowie die Dreifachinteraktion an der auch die Testlänge beteiligt ist, dürfte wie in Studie 3a darauf zurückzuführen sein, dass die Unterschiede zwischen den einzelnen Selektionsverfahren mit zunehmender Kommunalität des Kriteriums ansteigen, wie man an der Skalierung der Ordinaten in Abbildung 42 (S. 119) und Abbildung 43 (S. 120) erkennt. Bei sehr umfangreichen Itempool

scheint das MAXVAL-Verfahren bei einer geringen Kommunalität zu deutlich weniger validen Tests zu führen als die Selektion anhand der Itemvalidität und der Trennschärfe, wenn nur ein kleiner Teil der Items in den Test aufgenommen wird. Dieser Effekt war in Studie 3a wesentlich schwächer. Er tritt jedoch nur bei kleinen Stichprobenumfängen auf, wie eine Analyse der signifikanten Vierfachinteraktion der Faktoren Selektionsmethode, Testlänge, Kommunalität des Kriteriums und Stichprobenumfang zeigt (vgl. Abbildung 48 auf S. 125).

Die signifikante Interaktion der Selektionsmethode mit der Anzahl der Faktoren (vgl. Abbildung 44 auf S. 121 und Abbildung 45 auf S. 122) dürfte im Gegensatz zu Studie 3a darauf zurückzuführen sein, dass der Unterschied zwischen der Selektion anhand der Trennschärfe und der Selektion anhand der Itemvalidität deutlicher hervortritt, je *weniger* latente Faktoren der Itempool hat. Wie bereits erwähnt ist das MAXVAL-Verfahren bei umfangreichen Itempools – so wie in Studie 3a – nicht besser als die Selektion anhand der Itemvalidität, wenn nur ein kleiner Teil der Items in den Test aufgenommen. Allerdings scheint dieser Effekt in Studie 3b kaum von der Anzahl der latenten Dimensionen des Itempools abzuhängen.

Im Gegensatz zu Studie 3a hängt der Vergleich der zufälligen Auswahl mit der Methode zur Maximierung der internen Konsistenz von der Anzahl der latenten Faktoren ab. Bei vielen latenten Dimensionen führt die Auswahl anhand der internen Konsistenz zu deutlich valideren Skalen als die zufällige Auswahl. Bei wenig latenten Dimensionen resultieren bei zufälliger Auswahl dagegen validere Skalen als bei der Optimierung der internen Konsistenz, wenn nicht nur wenige Items in den Test aufgenommen werden.

Die Interaktion der Selektionsmethode mit dem Stichprobenumfang (Abbildung 40 auf S. 117 und Abbildung 41 auf S. 118) dürfte, so wie in Studie 3a, vor allem darauf zurückzuführen sein, dass das MAXVAL-Verfahren bei geringem Stichprobenumfang in der Population schlechter abschneidet, wenn nicht ein Großteil der Items in den Test aufgenommen wird. Dieser Effekt ist bei umfangreichen Itempools besonders markant.

Abbildung 40: Interaktion des Stichprobenumfangs mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Population (Fisher Z-transformiert) in Studie 3b (getrennte Darstellung je nach Umfang des Itempools).

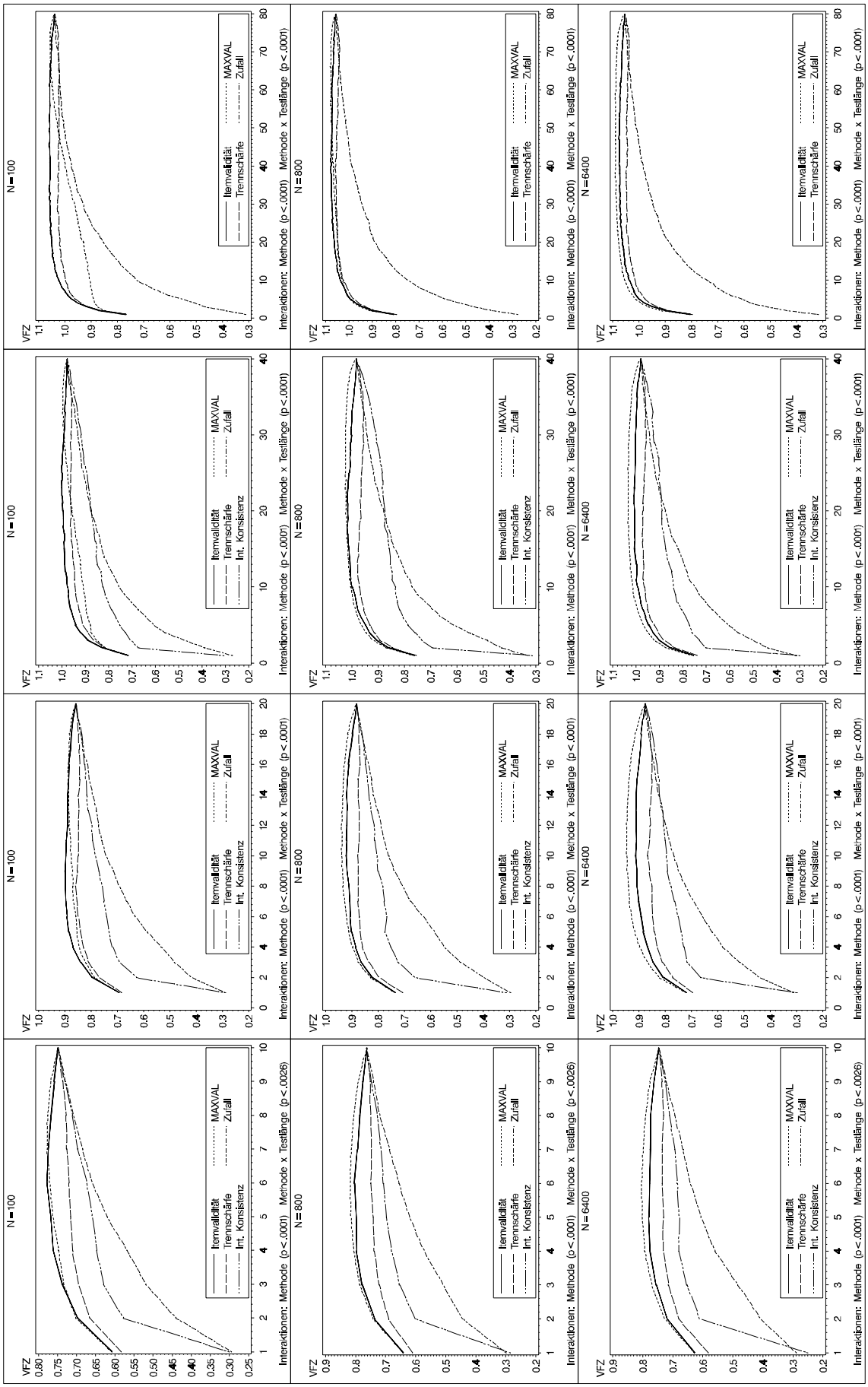
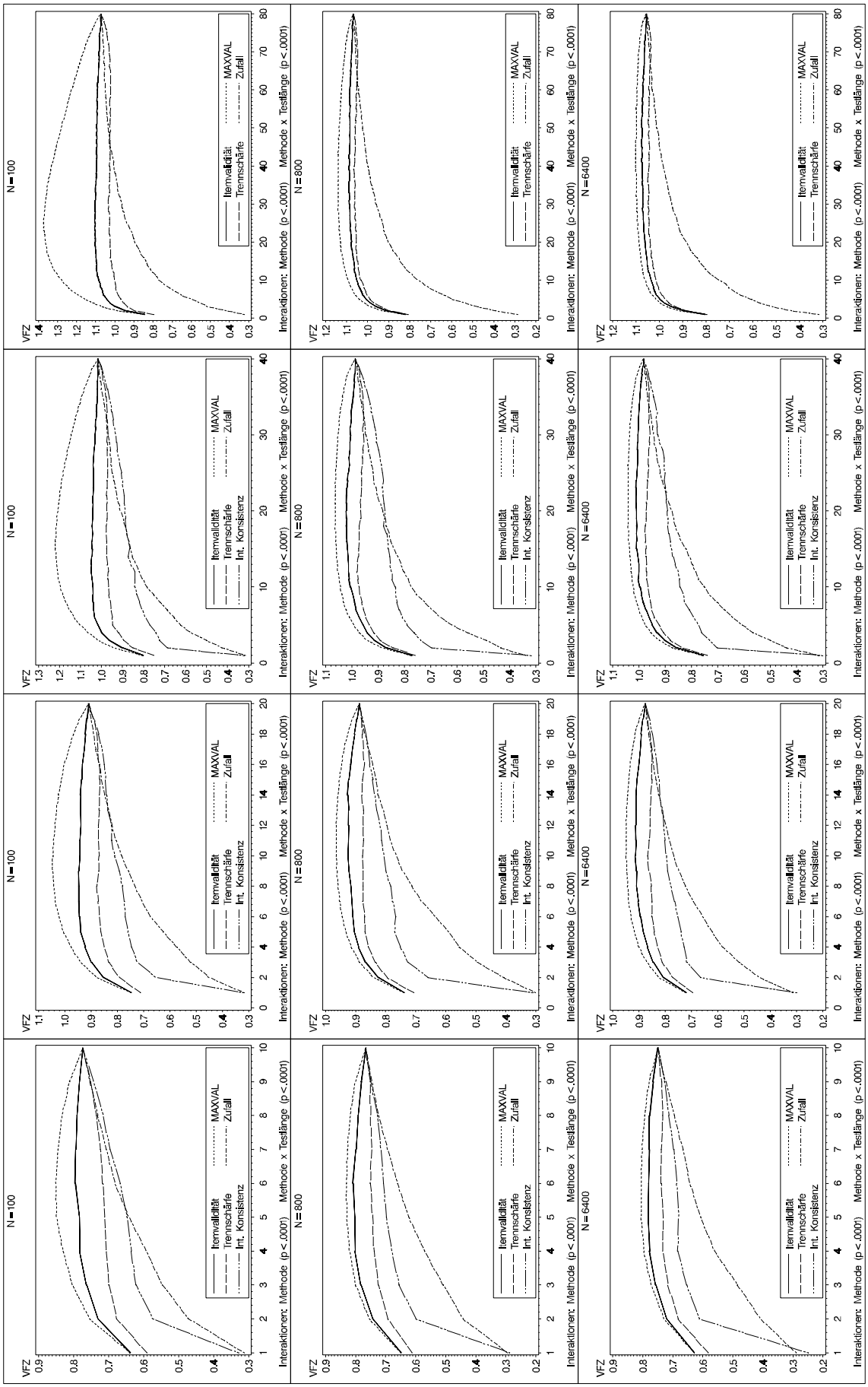


Abbildung 41: Interaktion des Stichprobenumfangs mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Stichprobe (Fisher Z-transformiert) in Studie 3b (getrennte Darstellung je nach Umfang des Itempools).



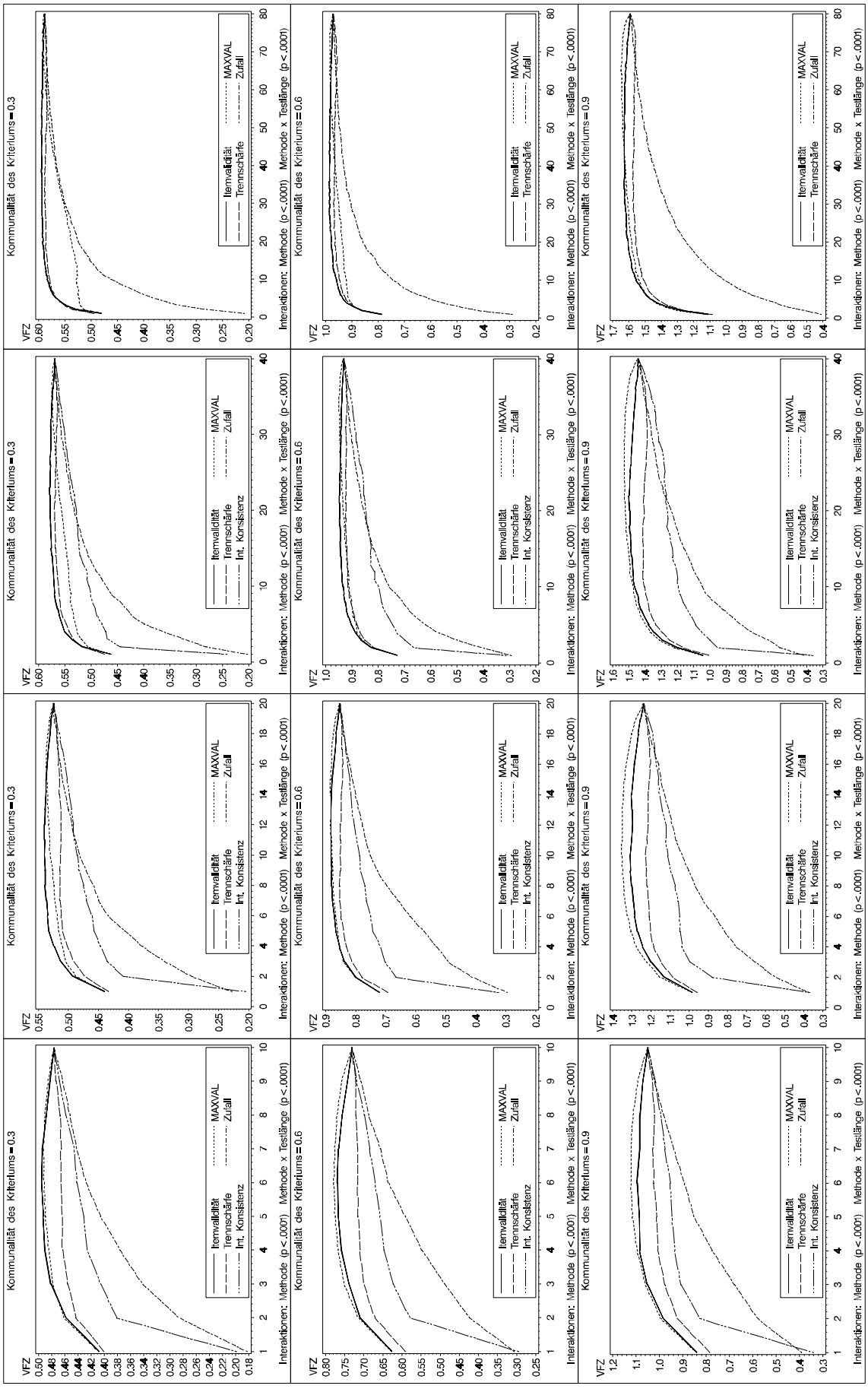


Abbildung 42: Interaktion der Kommunalität des Kriteriums mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Population (Fisher Z-transformiert) in Studie 3b (getrennte Darstellung je nach Umfang des Itempools).

Abbildung 43: Interaktion der Kommunalität des Kriteriums mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Stichprobe (Fisher Z-transformiert) in Studie 3b (getrennte Darstellung je nach Umfang des Itempools).

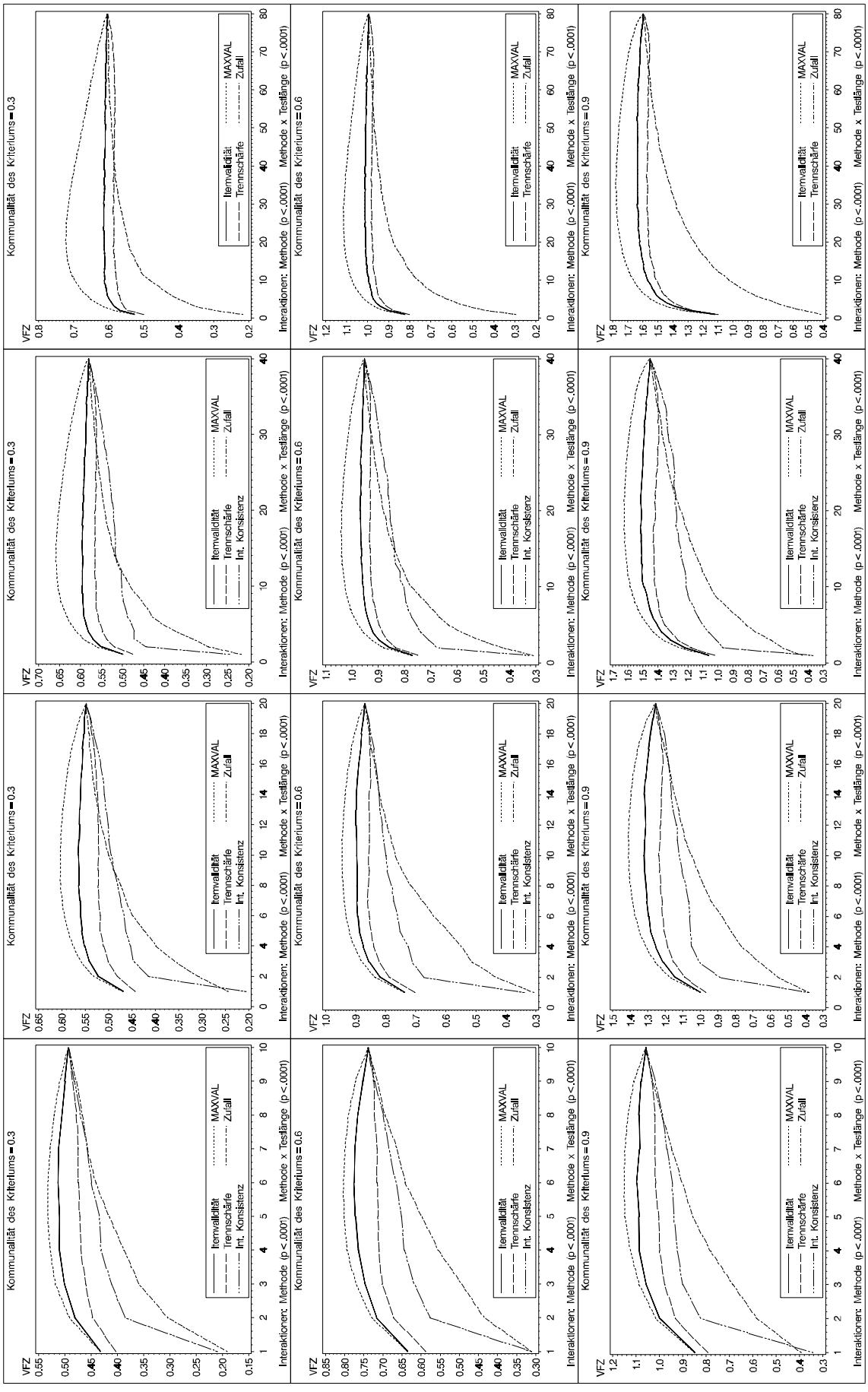


Abbildung 44: Interaktion der Dimensionalität des Itempools mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Population (Fisher Z-transformiert) in Studie 3b (getrennte Darstellung je nach Umfang des Itempools).

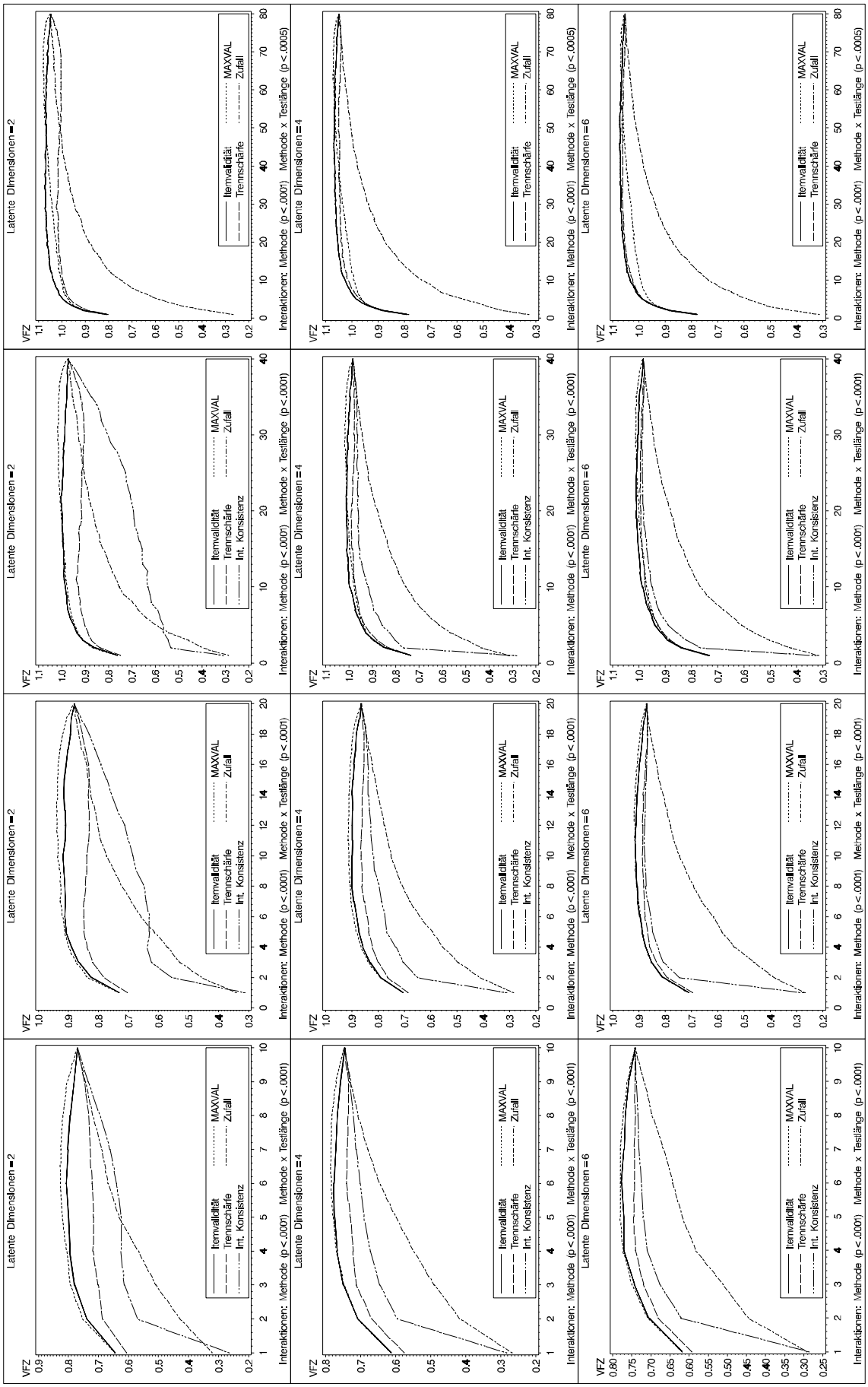


Abbildung 45: Interaktion der Dimensionalität des Itempools mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Validität in der Stichprobe (Fisher Z-transformiert) in Studie 3b (getrennte Darstellung je nach Umfang des Itempools).

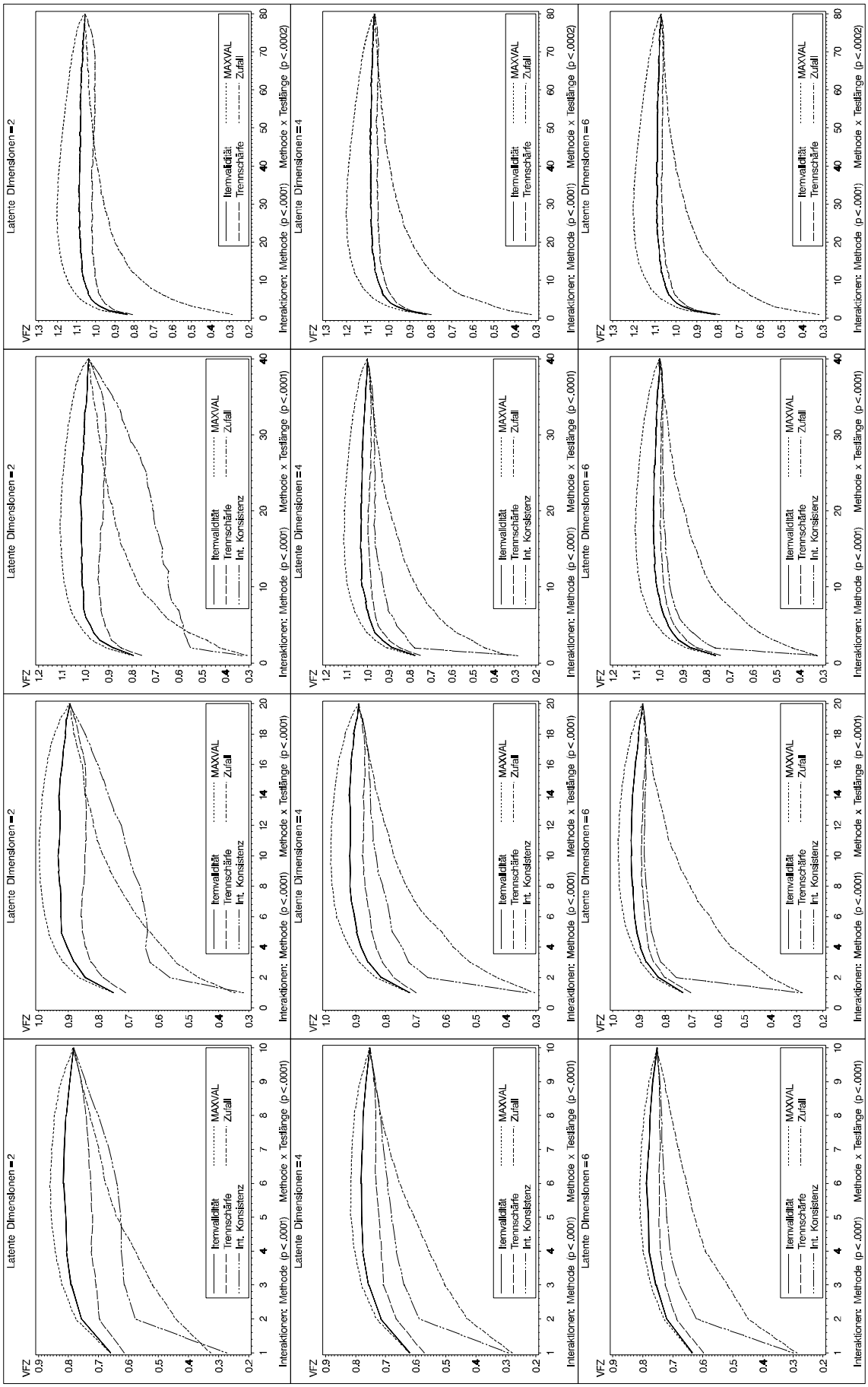


Abbildung 46: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und der Dimensionalität des Itempools bei der Vorhersage der Validität in der Population in Studie 3b (Testlänge 40)

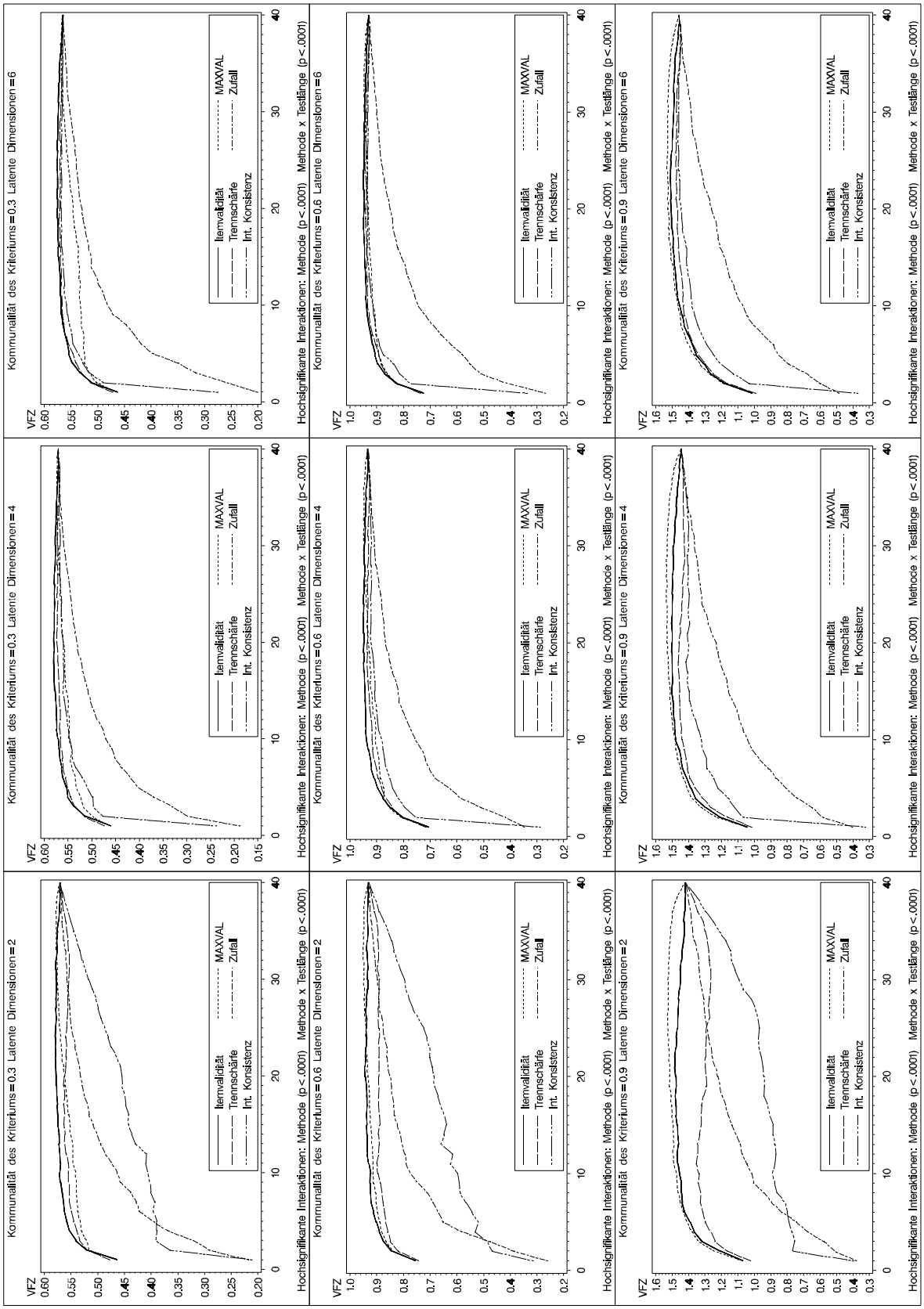


Abbildung 47: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und der Dimensionalität des Itempools bei der Vorhersage der Validität in der Population in Studie 3b (Testlänge 80)

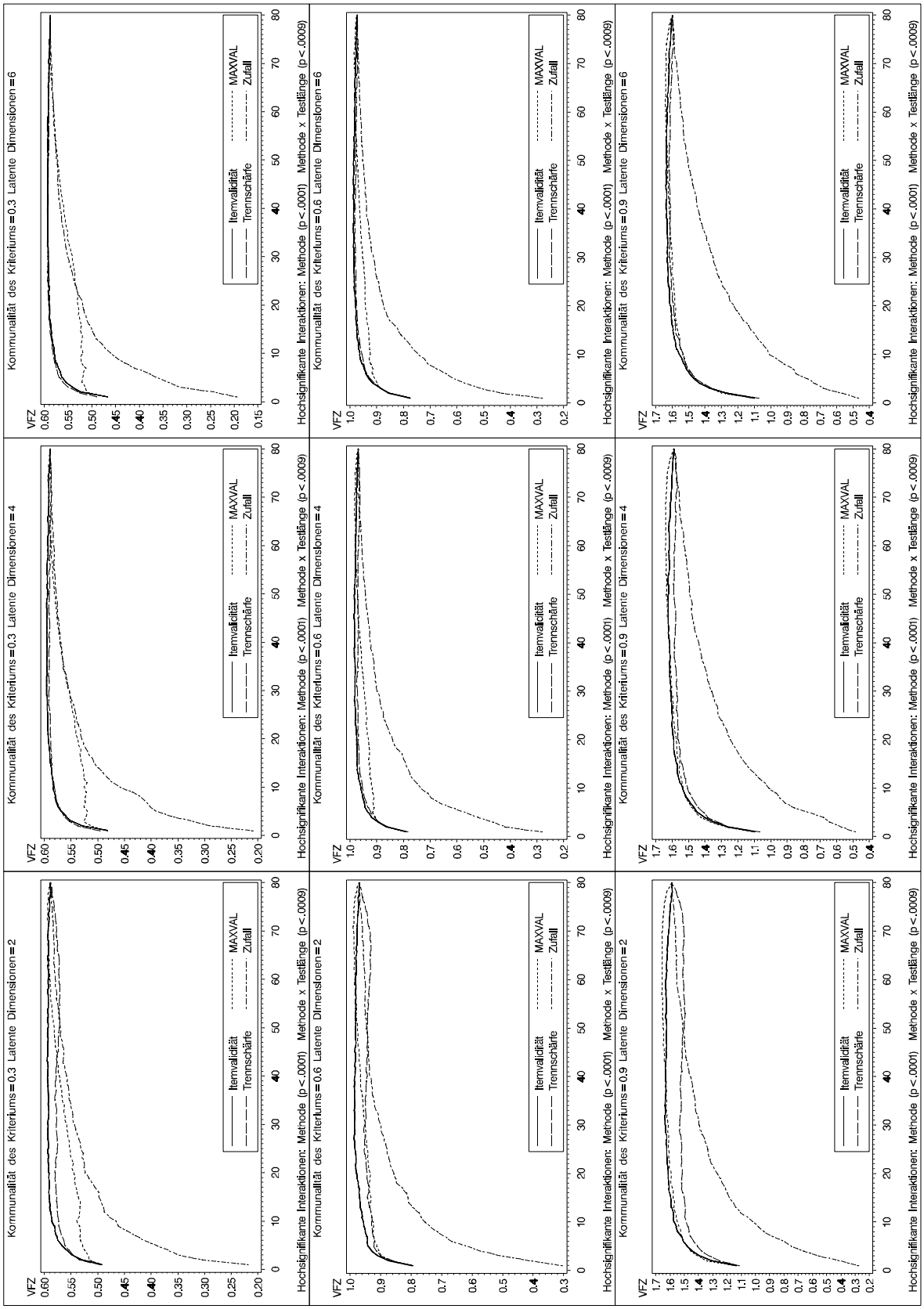
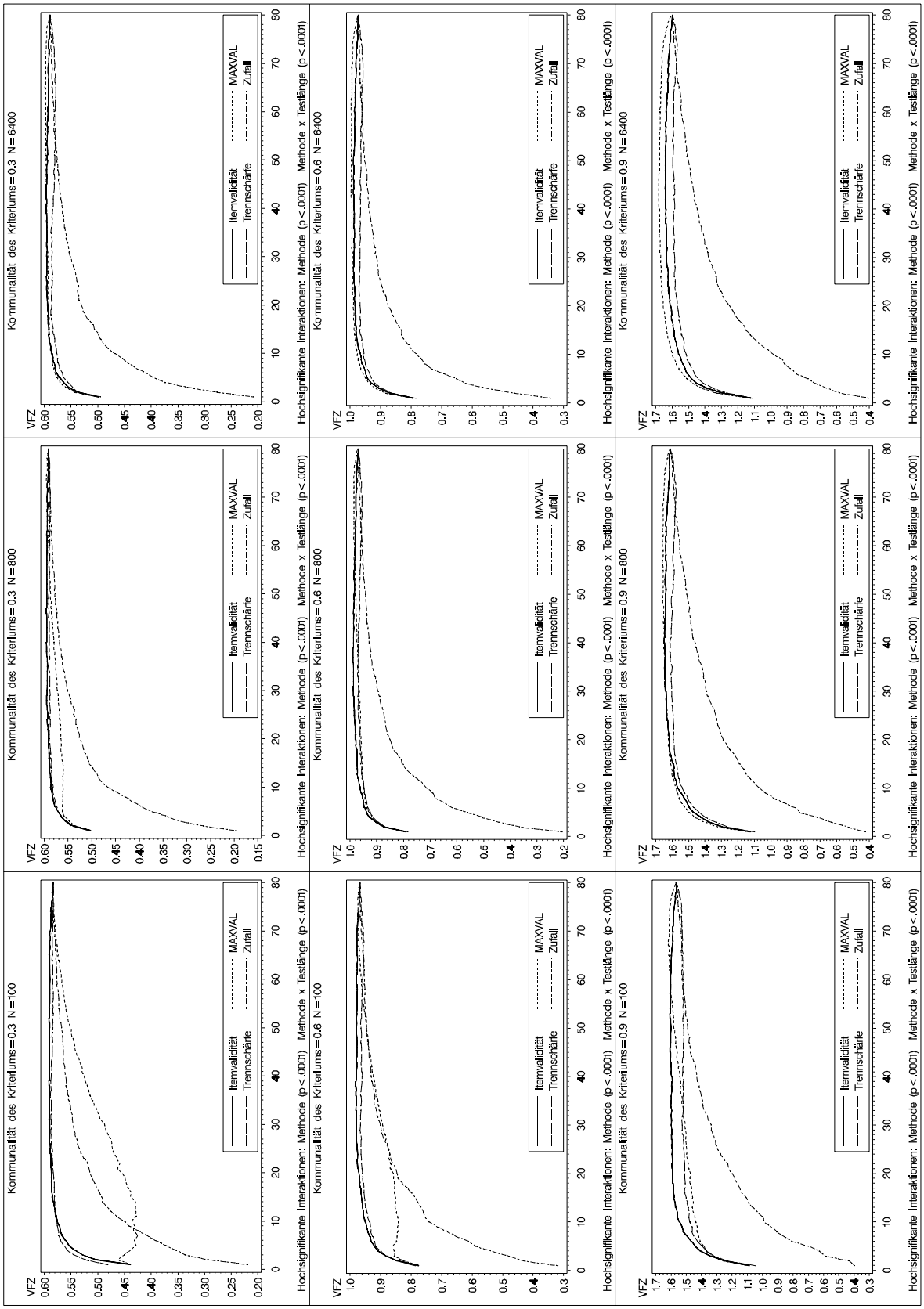


Abbildung 48: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und des Stichprobenumfangs bei der Vorhersage der Validität in der Population in Studie 3a (Testlänge 80)



Reliabilität

Haupteffekt der Methode und deren Interaktion mit der Testlänge

Der Vergleich der Reliabilität der mit den verschiedenen Selektionsmethoden ausgewählten Tests führte zu ähnlichen Ergebnissen wie bei Studie 3a. Allerdings ergaben sich in Studie 3b deutlich geringere Unterschiede zwischen den Selektionsverfahren (vgl. Abbildung 49). Nur die Reliabilität der mit dem MAXVAL-Verfahren ausgewählten Skalen sowie die der zufällig ausgewählten Verfahren unterscheidet sich in substantieller von den anderen Selektionsverfahren, wobei beim MAXVAL-Verfahren bei wenig umfangreichen Itempools deutlich reliablere Skalen resultieren als in Studie 3a. Bei umfangreichen Itempools unterscheidet sich die Reliabilität der mit dem MAXVAL-Verfahren ausgewählten Skalen jedoch nur dann substantiell von der zufälligen Auswahl, wenn nur ein geringer Teil der Items in den Test aufgenommen wird.

Abbildung 49: Reliabilität der verschiedenen Selektionsverfahren in Studie 3b

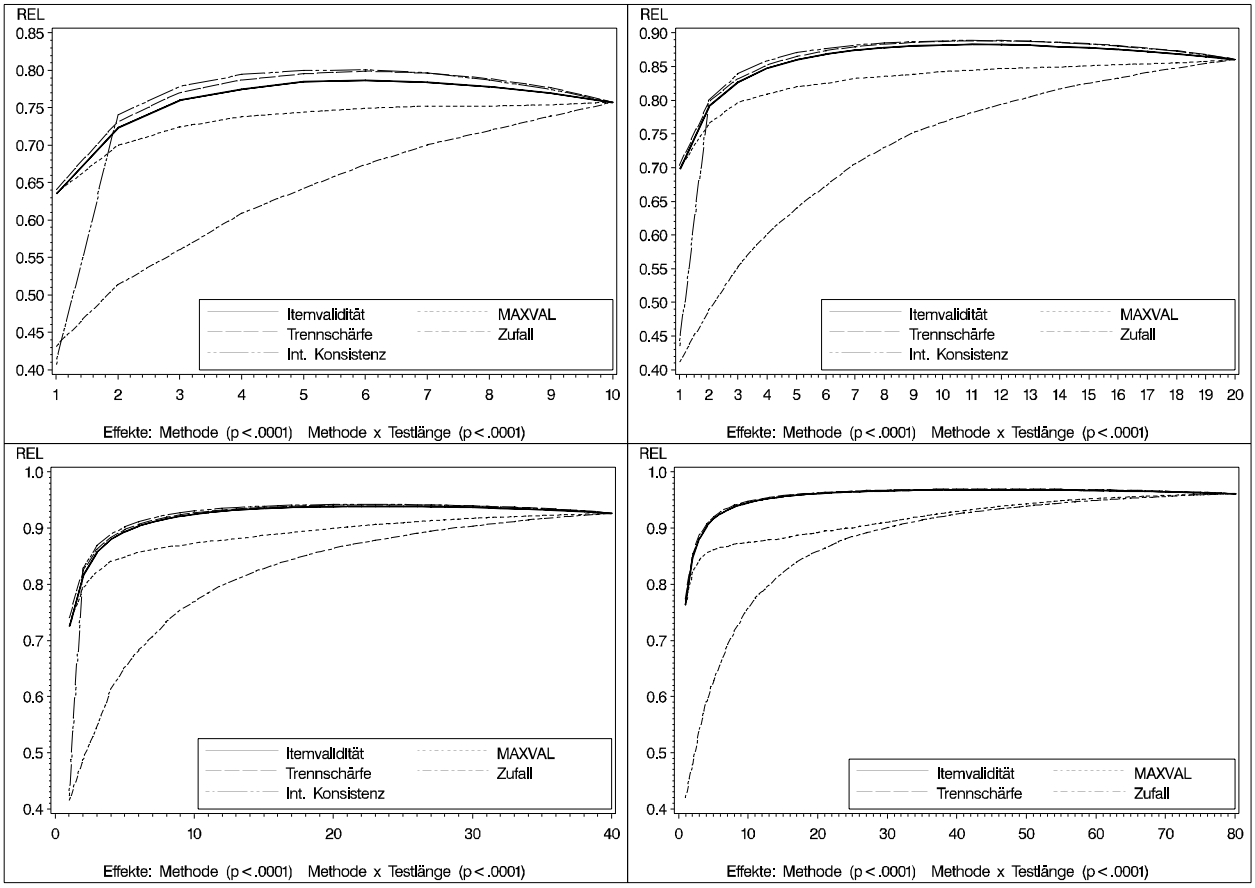


Abbildung 50: Interaktion des Stichprobenumfangs mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Reliabilität in Studie 3b (getrennte Darstellung je nach Umfang des Itempools).

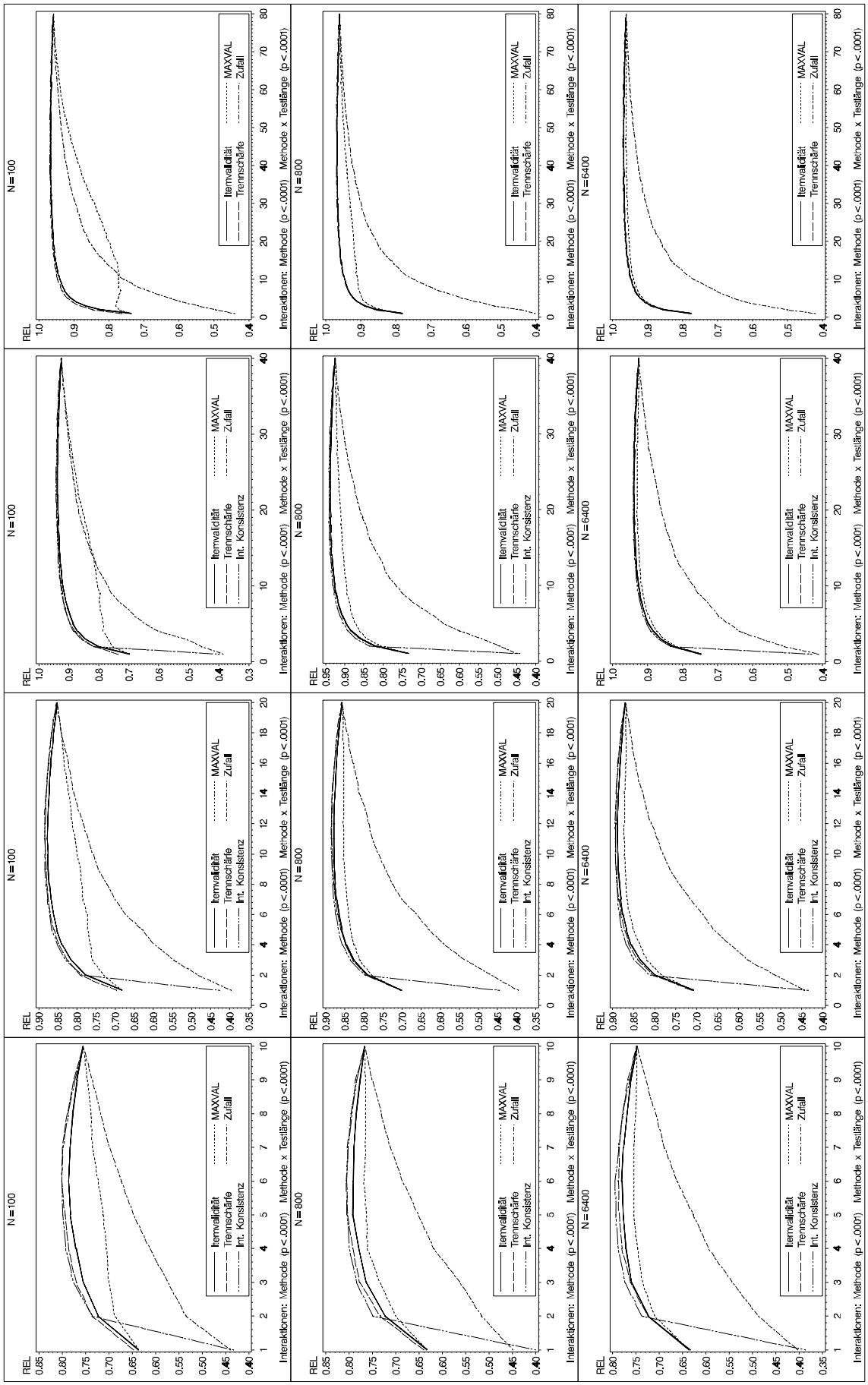


Abbildung 51: Interaktion der Kommunalität des Kriteriums mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Reliabilität in Studie 3b (getrennte Darstellung je nach Umfang des Itempools).

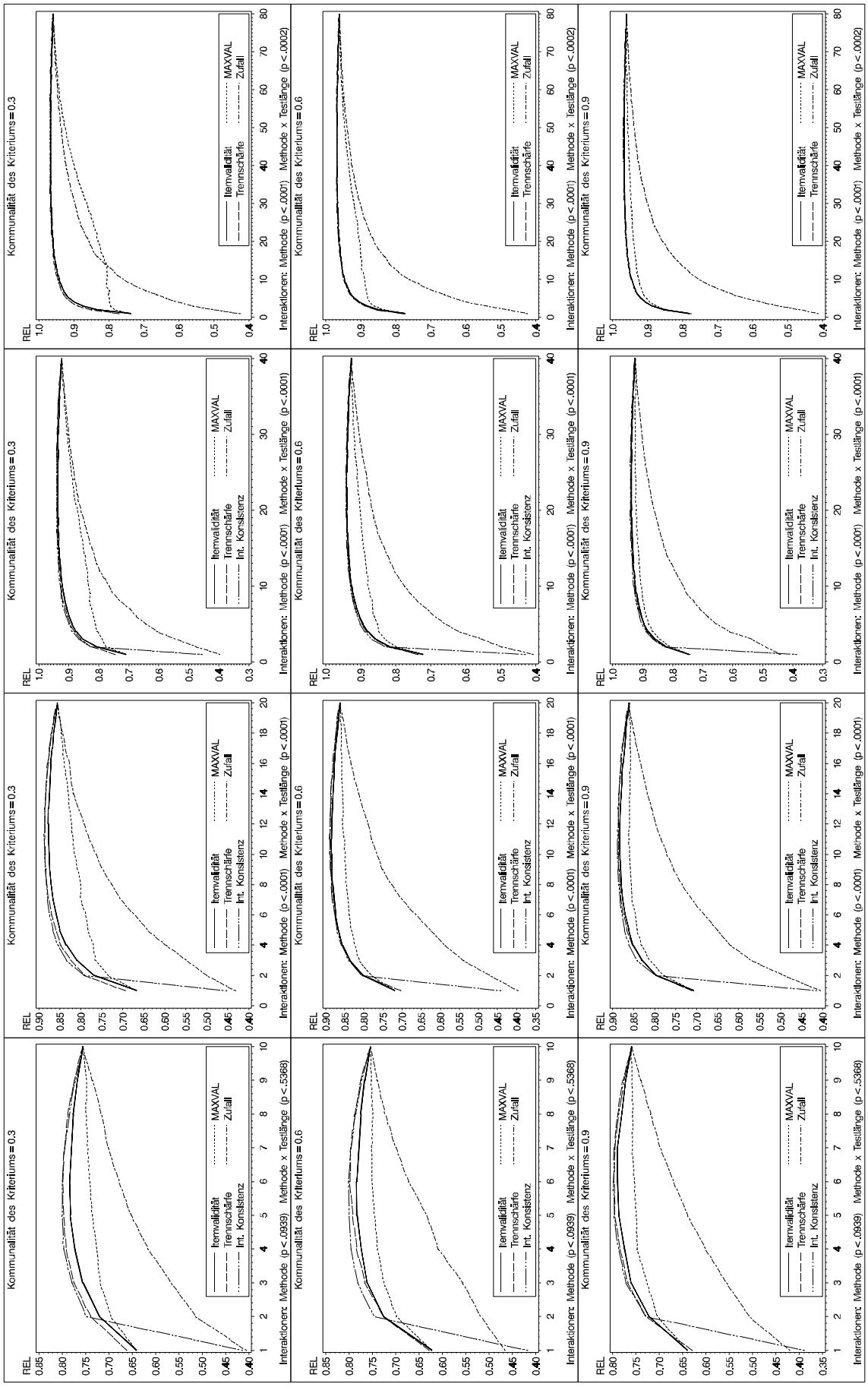


Abbildung 52: Interaktion der Dimensionalität des Itempools mit der Selektionsmethode (und der Testlänge) bei der Vorhersage der Reliabilität in Studie 3b (getrennte Darstellung je nach Umfang des Itempools).

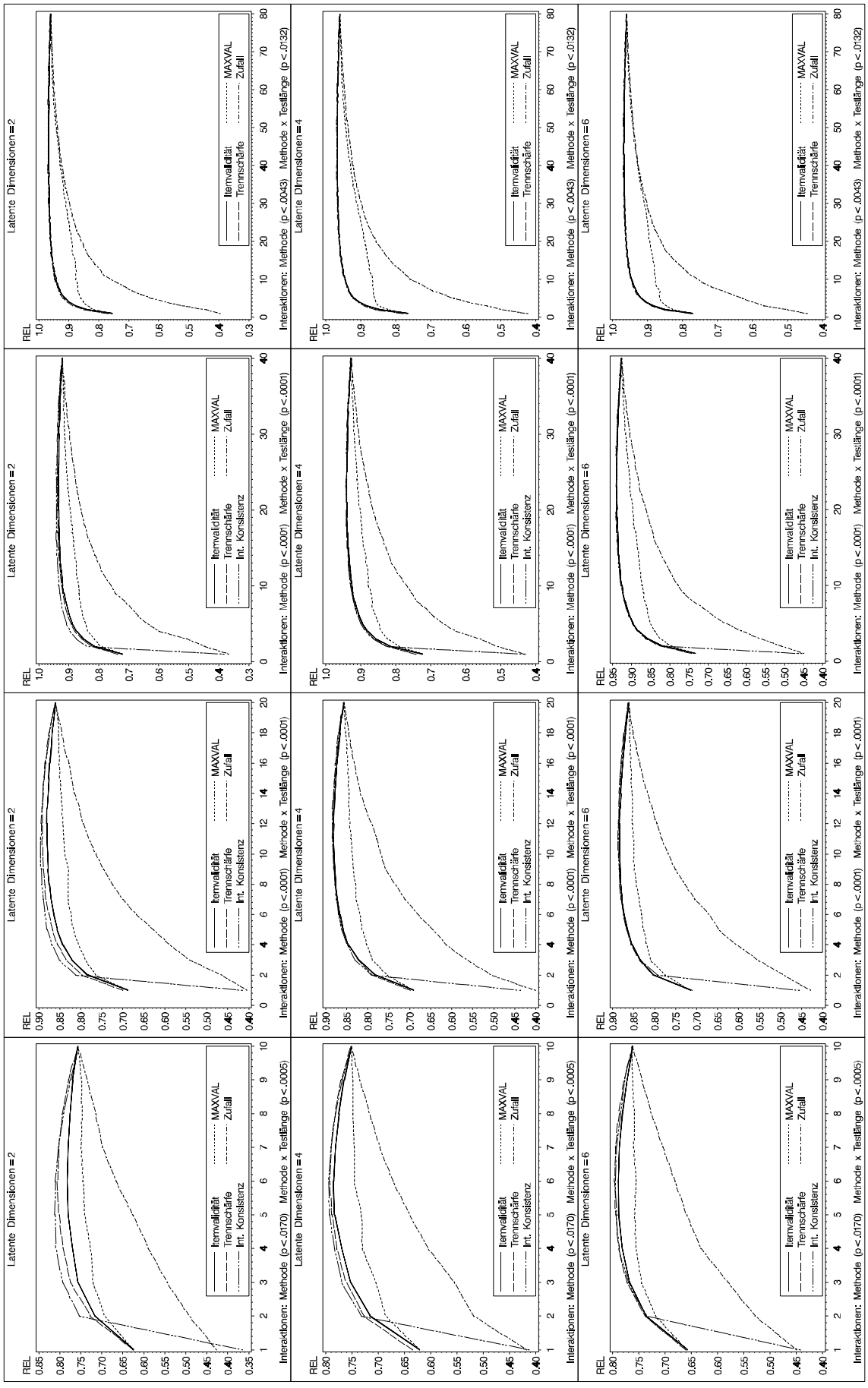


Abbildung 53: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Reliabilität in Studie 3b (Testlänge 10)

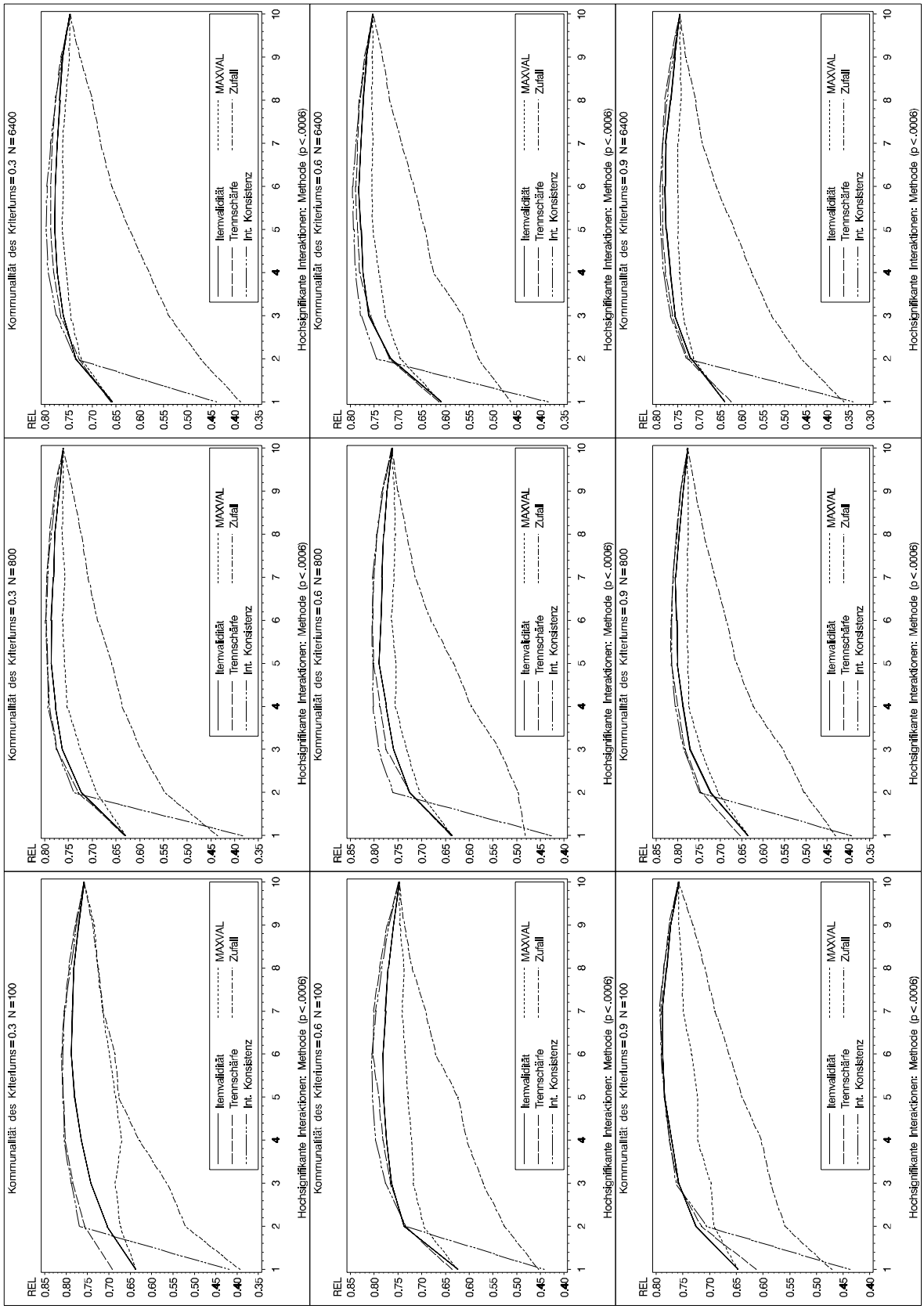


Abbildung 54: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Reliabilität in Studie 3b (Testlänge 20)

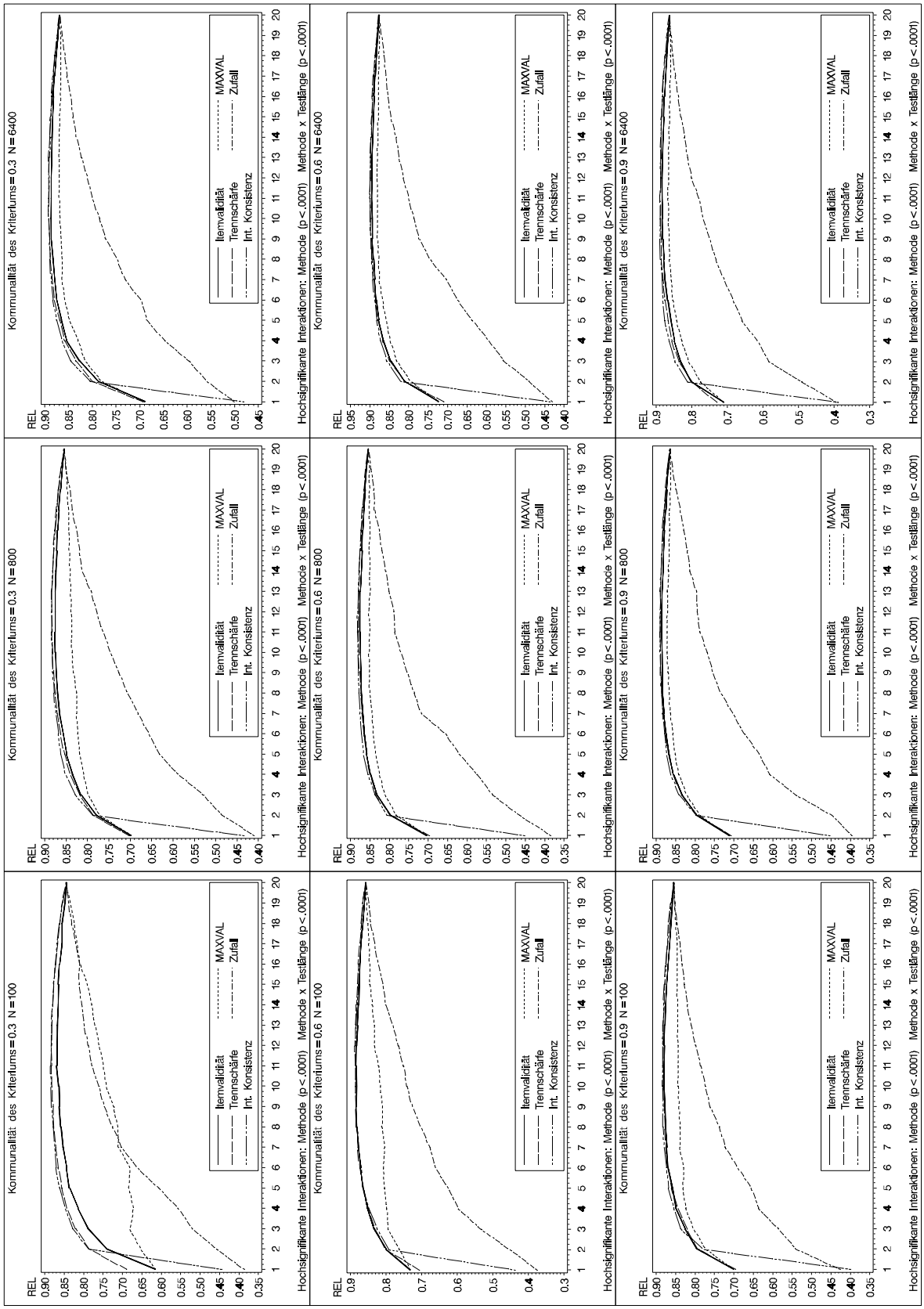


Abbildung 55: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Reliabilität in Studie 3b (Testlänge 40)

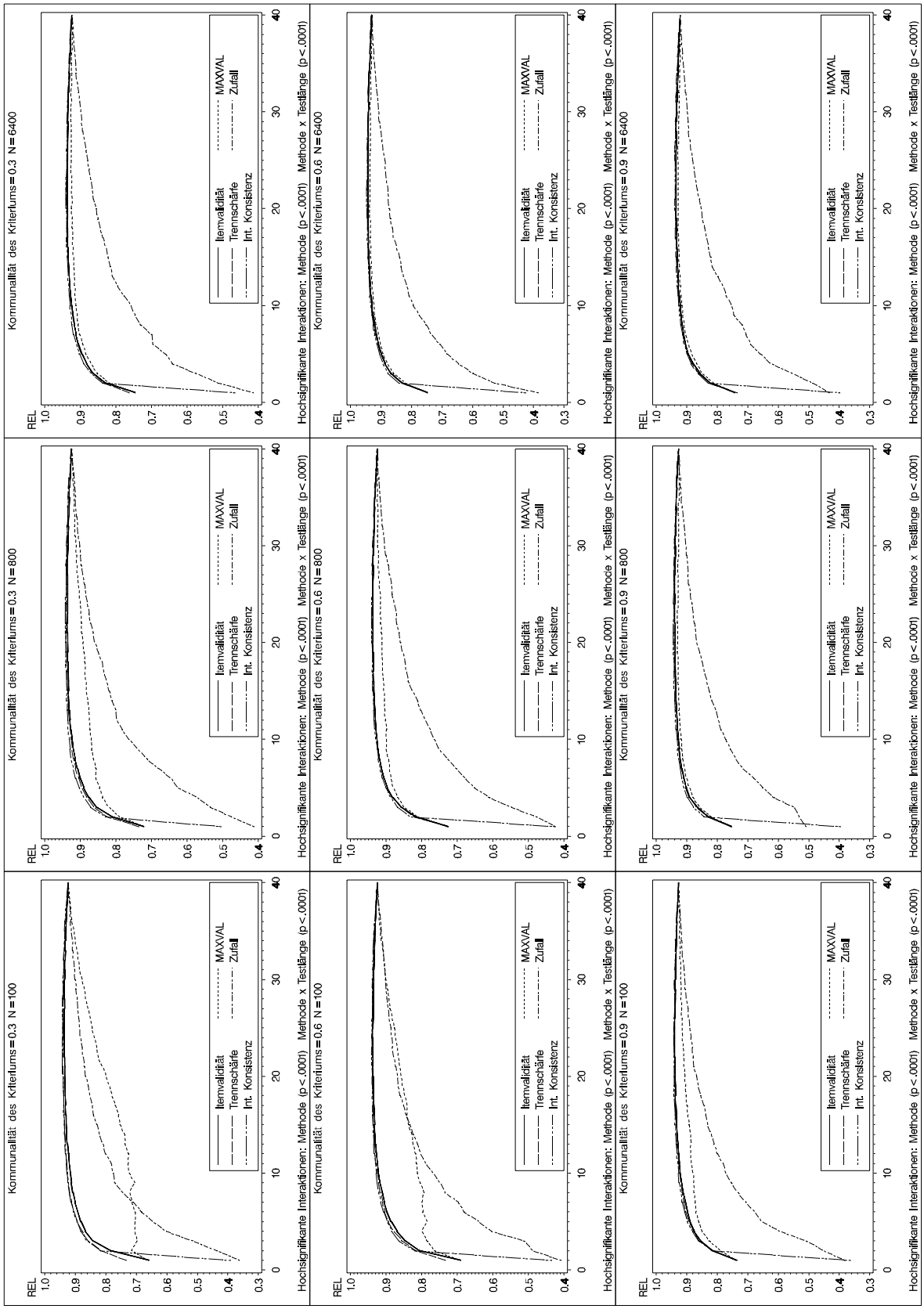
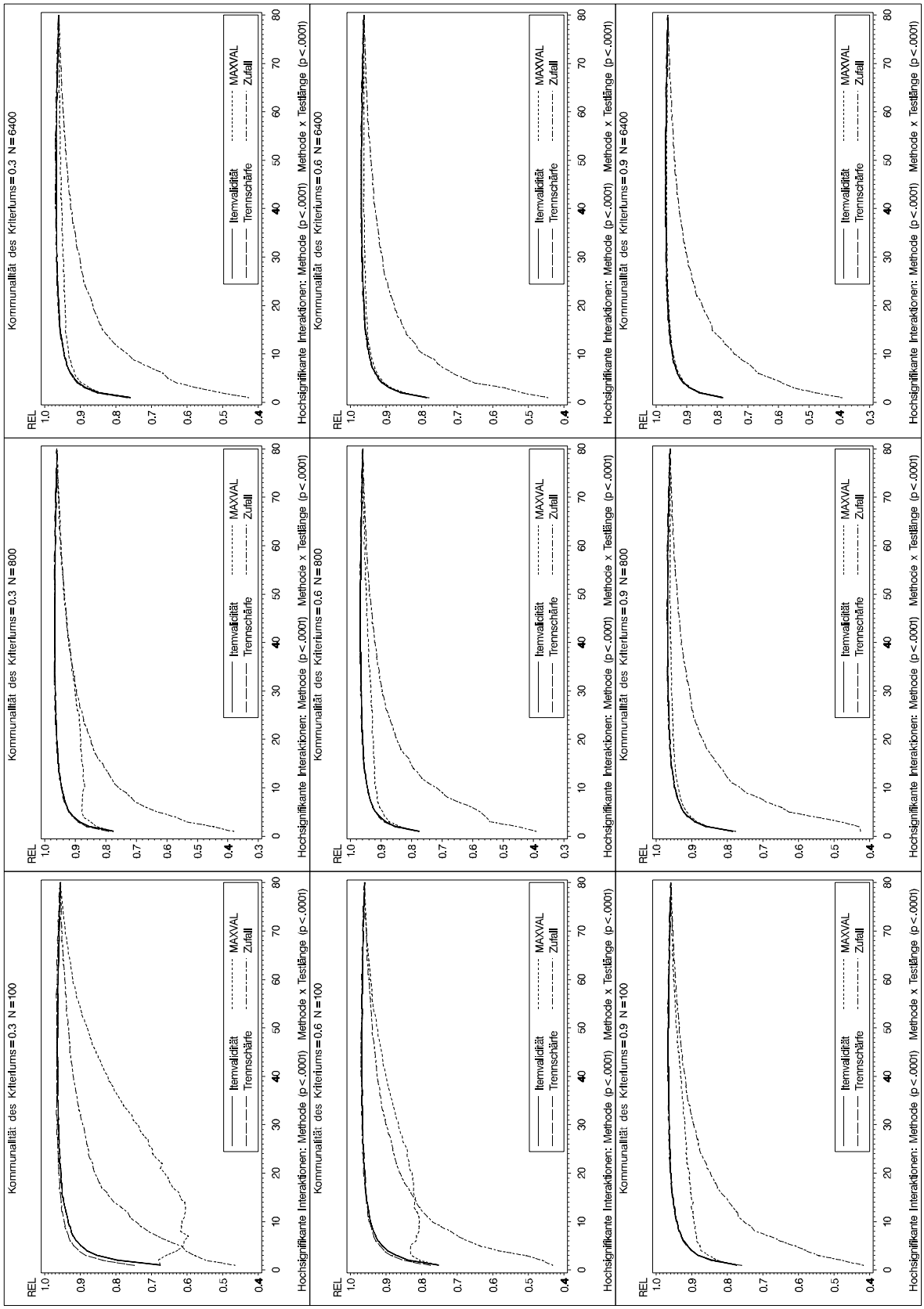


Abbildung 56: Interaktionen der Selektionsmethode mit der Kommunalität des Kriteriums und dem Stichprobenumfang bei der Vorhersage der Reliabilität in Studie 3b (Testlänge 80)



Interaktionen der anderen Haupteffekte mit der Selektionsmethode

So wie in Studie 3a gibt es bei umfangreicheren Itempools eine Interaktion der Kommunalität des Kriteriums mit der Selektionsmethode sowie eine Dreifachinteraktion an der zusätzlich die Testlänge beteiligt ist (vgl. Abbildung 51 auf S. 128) und eine Vierfachinteraktion unter Einschluss des Stichprobenumfangs (vgl. Abbildung 53, S. 130 bis Abbildung 56, S. 133). Dies scheint auch in Studie 3b darauf zurückzuführen zu sein, dass das MAXVAL-Verfahren v.a. bei geringer Kommunalität des Kriteriums und geringem Stichprobenumfang deutlich schlechter als die anderen Selektionsverfahren abschneidet.

Obwohl die Interaktion der Selektionsmethode mit der Anzahl der latenten Faktoren (vgl. Abbildung 52 auf S. 129) sowie die Dreifachinteraktion unter Einschluss der Testlänge bei Itempools mit mittlerem Umfang höchstsignifikant sind, ist der Moderatoreffekt der Dimensionalität des Itempools in Studie 3b deutlich geringer als in Studie 3a. Im Gegensatz zu Studie 3a scheinen die Unterschiede zwischen der Selektion anhand der Trennschärfe, der Selektion anhand der Itemvalidität und der Optimierung der internen Konsistenz mit zunehmender Anzahl der Faktoren kleiner zu werden.

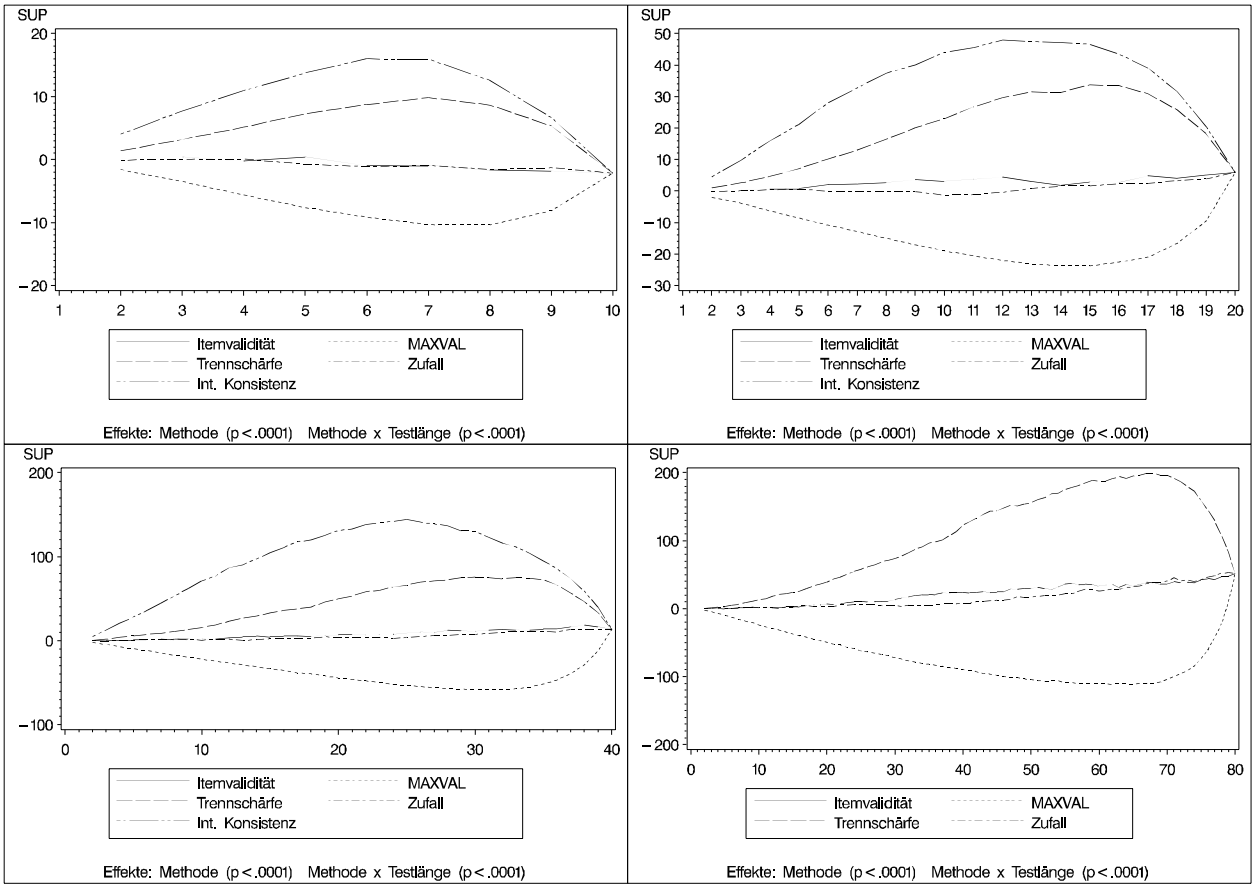
So wie in Studie 3a gibt es bei umfangreicheren Itempools eine signifikante Interaktion der Selektionsmethode mit dem Stichprobenumfang (vgl. Abbildung 50 auf S. 127). Sie resultiert daher, dass das MAXVAL-Verfahren bei umfangreicheren Itempools nur bei einem geringen Stichprobenumfang zu weniger reliablen Skalen führt als die anderen Selektionsverfahren. Dieser Effekt ist bei umfangreicheren Itempools besonders ausgeprägt. Hier resultieren selbst bei zufälliger Auswahl der Items Skalen mit höherer Reliabilität als beim MAXVAL-Verfahren, wenn nicht nur ein kleiner Teil der Items in den Test aufgenommen wird.

Suppression

Haupteffekt

So wie in Studie 3a ist bei den Verfahren, die eine Homogenisierung des Itempools anstreben, Fehlerredundanz zu beobachten, während beim MAXVAL-Verfahren Fehlersuppression vorliegt (vgl. Abbildung 57). Im Gegensatz zu Studie 3a hat sich das Ausmaß der Fehlerredundanz bei Selektion anhand der Trennschärfe sowie bei der Optimierung der internen Konsistenz deutlich verringert. Die Fehlersuppression des MAXVAL-Verfahrens ist jedoch kaum geringer als in Studie 3a und entspricht in der Größenordnung (im Durchschnitt) nun ziemlich genau der Fehlerredundanz, die bei Selektion anhand der Trennschärfe zu beobachten ist.

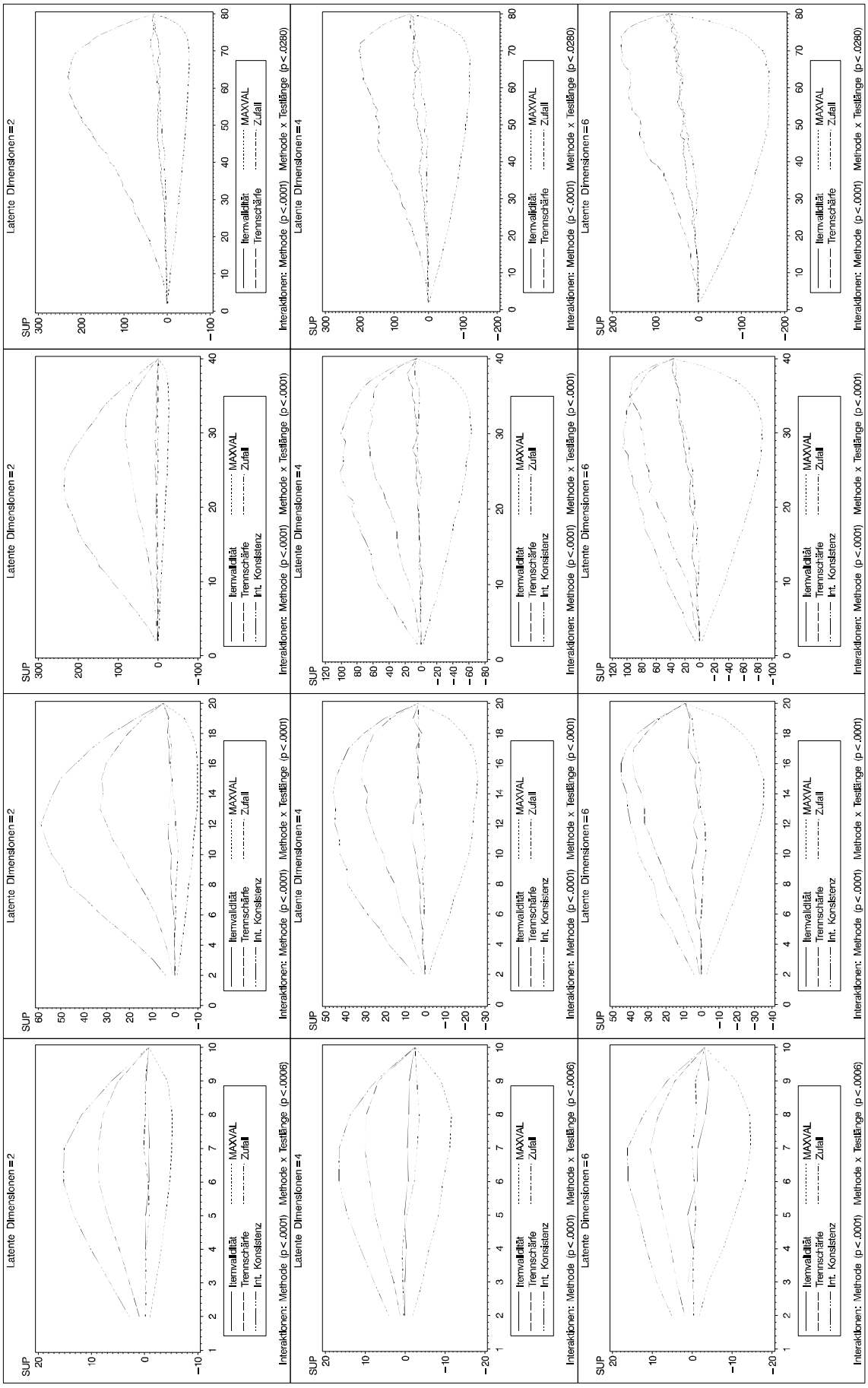
Abbildung 57: Kovarianz der Residuen bei Herauspartialisierung der wahren Kriteriumswerte



Interaktionen

So wie in Studie 3a ist eine signifikante Interaktion der Selektionsmethode mit der Anzahl der latenten Dimensionen zu beobachten (vgl. Abbildung 58). Die Fehlersuppression des MAXVAL-Verfahrens nimmt so wie in Studie 3a mit der Anzahl der latenten Dimensionen zu. Die Fehlerredundanz bei Selektion anhand der Trennschärfe ist nun jedoch weitgehend unabhängig von der Anzahl der latenten Dimensionen. Die Fehlerredundanz bei Optimierung von Cronbachs α wird jedoch im Gegensatz zu Studie 3a bei umfangreicheren Itempools mit zunehmender Anzahl der latenten Dimensionen eher geringer.

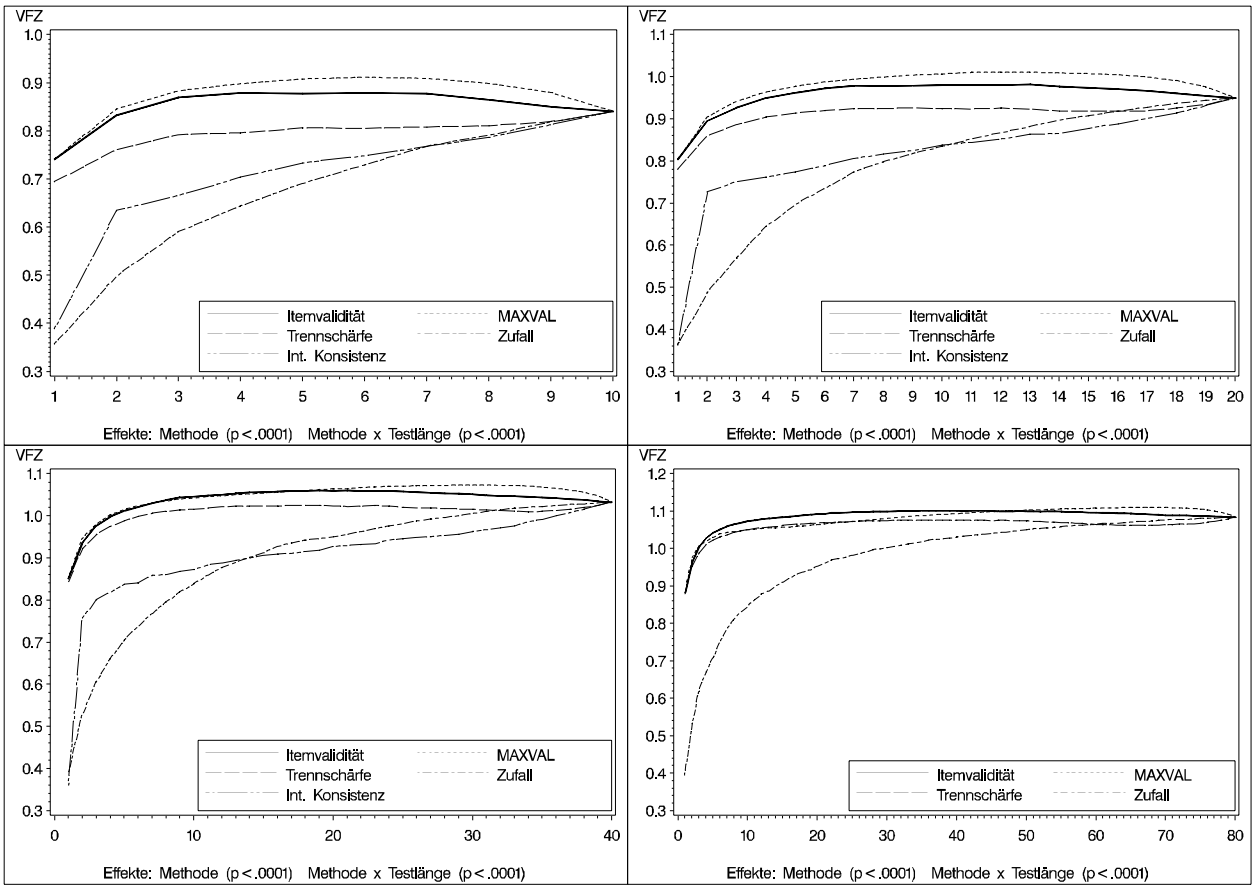
Abbildung 58: Interaktionen der Selektionsmethode mit der Anzahl der Dimensionen des Itempools bei der Vorhersage der Kovarianz der Residuen bei Herauspartialisierung der wahren Kriteriumswerte in Studie 3b (getrennt mach Umfang des Itempools)



5.2.3.c Studie 3c

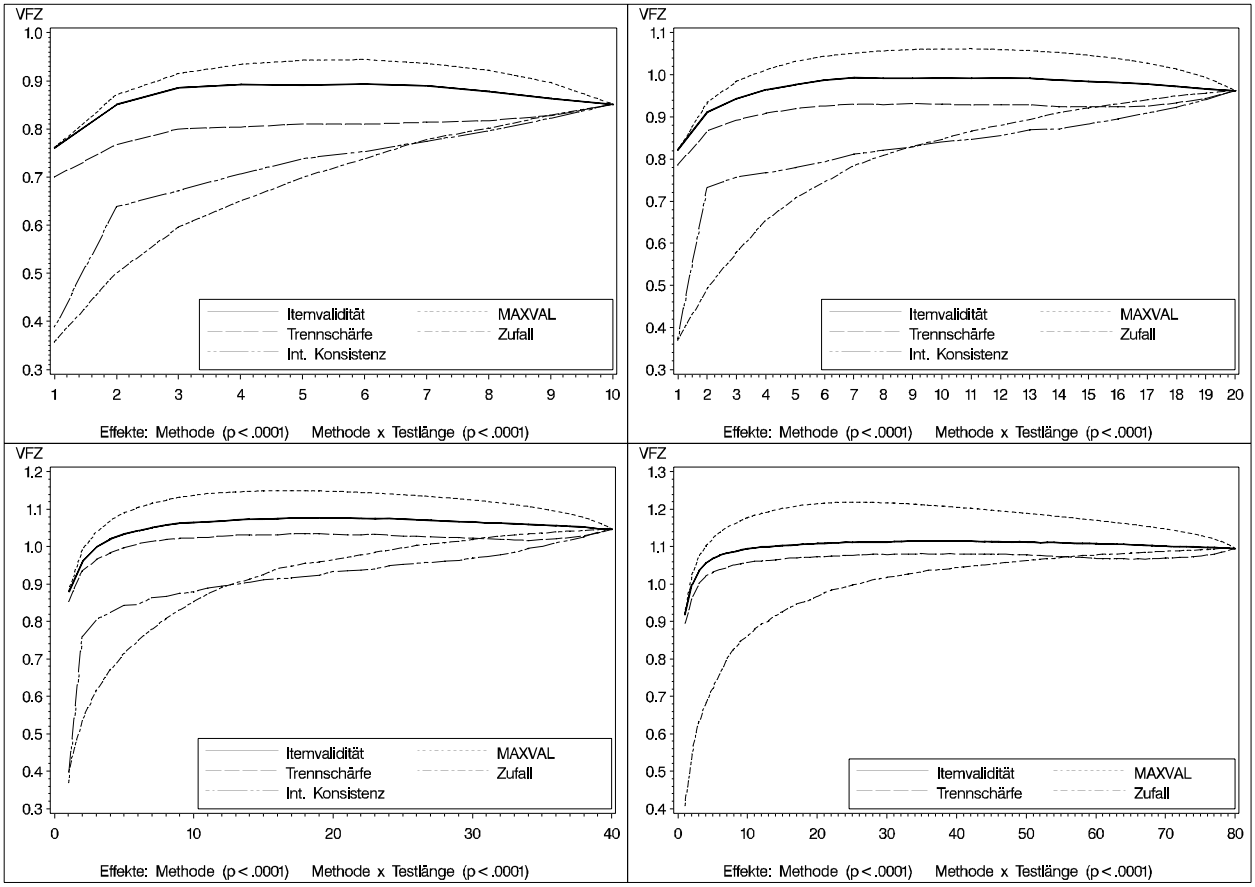
Studie 3c unterscheidet sich von Studie 3b und Studie 3a dadurch, dass die spezifische Varianz der Items im Mittel nicht die Hälfte, sondern lediglich ein Drittel der Gesamtvarianz der Items ausmacht. So wie in Studie 3b verteilt sich die gemeinsame Varianz je zur Hälfte auf den validen Faktor und auf die gemeinsamen Fehlerfaktoren. Das MAXVAL-Verfahren schneidet in Studie 3c geringfügig besser ab als in Studie 3b (vgl. Abbildung 59 und Abbildung 60). Die Selektion anhand der Trennschärfe war zwar auch in Studie 3b etwas schlechter als die zufällige Auswahl, wenn nur wenige Items eines umfangreichen Itempools nicht in den Test aufgenommen wurden. In Studie 3c tritt dieser Effekt jedoch etwas deutlicher hervor, auch wenn er nicht so stark ausgeprägt ist wie in Studie 3a.

Abbildung 59: Validität der verschiedenen Selektionsverfahren in Studie 3c (in der Population)



Ansonsten ist die durchschnittliche Validität der einzelnen Selektionsverfahren in der Stichprobe sowie der Population mit den Ergebnissen aus Studie 3b vergleichbar. Auch in Bezug auf die Reliabilität und die Kovarianz der Residuen bei Herauspartialisierung der wahren Kriteriumswerte ergaben sich weitgehend dieselben Ergebnisse wie in Studie 3b. Daher wird auf eine gesonderte Darstellung der weiteren Ergebnisse von Studie 3c verzichtet.

Abbildung 60: Validität der verschiedenen Selektionsverfahren in Studie 3c (in der Stichprobe)



5.3 Interpretation

Innerhalb und zwischen den dargestellten Simulationsstudien wurde eine Vielzahl von Faktoren variiert (vgl. Tabelle 3 auf S. 66 und Tabelle 4 auf S. 69). Zwischen den einzelnen Studien hat vor allem die Verteilung der Varianzkomponenten der Items variiert. Während in Studie 1 der valide Faktor nur einen geringen Anteil der Gesamtvarianz erklärt und der Rest gemeinsame Fehlervarianz ist, verteilt sich die Varianz der Items in Studie 2 im Mittel zu gleichen Teilen auf valide und spezifische Varianz. Nur in den Studien 3a, 3b und 3c kommen sowohl valide Varianz, gemeinsame Fehlervarianz und spezifische Varianz vor, wobei die Anteile der drei Varianzkomponenten zwischen den Studien variieren. In Studie 3a und Studie 3b ist im Durchschnitt die Hälfte der Itemvarianz spezifische Varianz, während der Anteil der spezifischen Varianz in Studie 3c nur ein Drittel ausmacht. In Studie 3b und 3c ist die Hälfte der gemeinsamen Varianz valide Varianz, während der Anteil der validen Varianz in Studie 3a mit der Anzahl der Faktoren abnimmt.

5.3.1 Validität

In der Stichprobe erreicht das MAXVAL-Verfahren in allen dargestellten Simulationsstudien unabhängig von der Verteilung der Varianzkomponenten die höchste Validität, gefolgt von der Selektion anhand der Itemvalidität. Bei der Optimierung von Cronbachs α und der Selektion anhand der Trennschärfe resultieren meist deutlich weniger valide Skalen. Es ist jedoch ein triviales Ergebnis, dass Verfahren, die Items anhand von Validitätsdaten in der Stichprobe auswählen, auch zu höheren Validitätskoeffizienten *in der Stichprobe* führen.

Für die Validität *in der Population* ist die Verteilung der Varianzkomponenten dagegen von entscheidender Bedeutung für die Frage, welches Selektionsverfahren zu valideren Tests führt. Je größer der Anteil der validen Varianz in Relation zur gemeinsamen Fehlervarianz der Items ist, desto besser sind die Trennschärfe und die interne Konsistenz als Selektionskriterien. Nur wenn – so wie in Studie 2 – gar keine gemeinsame Fehlervarianz vorhanden ist, scheint die Optimierung der internen Konsistenz und die Auswahl anhand der Trennschärfe nicht nur zu besonders reliablen, sondern auch zu besonders validen Skalen zu führen, während die Selektion nach dem MAXVAL-Verfahren hier grundsätzlich keine Vorteile bringt und zum Teil sogar zu Skalen mit deutlich geringerer Validität und Reliabilität führt. Wenn dagegen, so wie in Studie 1 und Studie 3a (bei mehr als 2 gemeinsamen Faktoren), der Anteil der gemeinsamen Fehlervarianz überwiegt, dann erscheint die Nutzbarmachung von Suppressionseffekten durch das MAXVAL-Verfahren sehr viel lohnender als die Selektion anhand der Trennschärfe. Dass das gute Abschneiden des MAXVAL-Verfahrens in diesen beiden Studien primär auf den hohen Anteil der gemeinsamen Fehlervarianz in Relation zur validen Varianz zurückzuführen ist und nicht auf die Anzahl der gemeinsamen Fehlerfaktoren, zeigt sich daran, dass bei Konstanthalten des Anteils der gemeinsamen Fehlervarianz in Studie 3b und 3c eine größere Anzahl gemeinsamer Fehlerfaktoren nicht mit einem wesentlich besseren Abschneiden des MAXVAL-Verfahrens einhergeht. Bei umfangreichen Itempools resultieren in Studie 3b und 3c bei Anwendung des MAXVAL-Verfahrens mit zunehmender Anzahl der gemeinsamen Fehlerfaktoren sogar weniger valide Skalen. Die Berücksichtigung von Suppressionseffekten ist demnach, wie erwartet, dann besonders lohnend, wenn sich viel gemeinsame Fehlervarianz auf möglichst wenige Faktoren verteilt.

Die Verringerung des spezifischen Varianzanteils der Items in Studie 3c führte nur zu einem unwesentlich besseren Abschneiden des MAXVAL-Verfahrens. Dass das MAXVAL-Verfahren in Studie 1, in der die Items gar keine spezifischen Fehleranteile hatten, so gut abgeschnitten hat, dürfte daher primär auf den hohen Anteil gemeinsamer Fehlervarianz zurückzuführen sein. Auch

sonst scheint der Anteil der spezifischen Varianz kein entscheidender Faktor für die beobachteten Validitätsunterschiede zwischen den Verfahren zu sein.

Bedeutsame Interaktionen der Testlänge mit den Selektionsverfahren waren vor allem darauf zurückzuführen, dass sich die Selektionsverfahren in den abhängigen Variablen kaum unterscheiden, wenn bereits ein Großteil der Items in den Test aufgenommen wurden. Dies ist ein trivialer Effekt, der darauf zurückzuführen ist, dass die von verschiedenen Selektionsverfahren ausgewählten Tests viele Items gemeinsam haben müssen, wenn nur ein geringer Anteil der Items nicht in den Test aufgenommen wird. Höhere Interaktionen der Testlänge mit Beteiligung von weiteren Faktoren waren in der Regel v.a. darauf zurückzuführen, dass das MAXVAL-Verfahren bei kurzen Tests in der Population schlechter abschnitt. In der Stichprobe waren dann jeweils deutlich überhöhte Schätzungen der Validität zu beobachten. Dass die Regression zur Mitte bei kurzen Tests deutlicher ausfällt, dürfte daran liegen, dass die Korrelation der Stichprobenfehler bei Ermittlung der Validität von längeren Tests wegen des „Item-Overlaps“ größer ist (vgl. [2.2-5] auf S. 38). Dieser Effekt konnte erwartungsgemäß vor allem bei umfangreichen Itempools beobachtet werden, da hier viel mehr Itemkombinationen zur Auswahl stehen. Es kommt allerdings nur bei geringer Kommunalität und geringerem Stichprobenumfang zu deutlich invalideren Skalen. Dies ist plausibel, da die Regression zur Mitte nur dann zu gravierenden Verzerrungen führen kann, wenn die Fehler bei Ermittlung der Validität anhand der Stichprobendaten groß ist (vgl. [2.2-4] auf S. 37). Ist der Fehler dagegen wegen eines großen Stichprobenumfangs gering, so reduziert sich die Regression zur Mitte auch bei kürzeren Skalen. Bei sehr großen Stichproben gehörte das MAXVAL-Verfahren auch in der Population immer zu den validesten Verfahren.

Es kommt jedoch weniger auf die absolute Höhe des Schätzfehlers an, sondern auf das Verhältnis zwischen dem Schätzfehler (der Validität) und der wahren Varianz der Validität bei den zur Disposition stehenden Itemkombinationen. Bei einer geringen Kommunalität des Kriteriums verringert sich die wahre Varianz der geschätzten Validitätskoeffizienten aufgrund von Verdünnungseffekten. Daher ist auch bei einer geringen Kommunalität des Kriteriums beim MAXVAL-Verfahren eine deutlich stärkere Regression zur Mitte zu beobachten. Ansonsten führt eine geringe Kommunalität des Kriteriums dazu, dass Unterschiede in der Validität der ausgewählten Skalen nivelliert werden. Wenn die Fehler bei der Schätzung der Kovarianzmatrix aufgrund einer sehr großen Personenstichprobe kaum ins Gewicht fallen, dann führt die Minderung der Validitätsunterschiede durch eine geringe Kommunalität des Kriteriums auch beim MAXVAL-Verfahrens kaum zu einer nennenswerten Regression zur Mitte. Der

entsprechende Interaktionsterm mit dem Stichprobenumfang erreichte jedoch nicht in allen Fällen die strengen Maßstäbe statistischer Signifikanz, die in dieser Studie angelegt wurden.

In Studie 1 konnte in der Population keine signifikante Interaktion der Testlänge mit der Selektionsmethode und dem Stichprobenumfang beobachtet werden. Hier war das Ausmaß der Regression zur Mitte bei allen Selektionsmethoden und unabhängig von der Testlänge etwa gleich stark ausgeprägt. Es ist denkbar, dass dies darauf zurückzuführen ist, dass die wahre Varianz in der Validität der verschiedenen Selektionsverfahren besonders groß ist.

Die Trennschärfe schneidet bei sehr kurzen Tests häufig nicht sehr viel schlechter ab als die beiden Verfahren, die auf Validitätsdaten basieren. Wenn dagegen nur wenige Items entfernt werden, so scheint die Trennschärfe meist nicht besser und zum Teil sogar etwas schlechter zu sein als die zufällige Auswahl[13]. Vor allem in Studie 1 und 3a war die Elimination weniger trennscharfer Items schlechter als die zufällige Selektion und zwar besonders bei umfangreichen Itempools. Dieser Effekt dürfte also vor allem dann auftreten, wenn der Anteil der gemeinsamen Fehlervarianz recht hoch ist und sich nicht auf zu viele Dimensionen verteilt. Auch bei einer hohen Kommunalität des Kriteriums scheint die Elimination weniger trennschwacher Items in der Regel nicht ratsam zu sein. Wenn der Anteil der gemeinsamen Fehlervarianz gering ist oder sich auf viele Faktoren verteilt, dann ist die Trennschärfe häufig nicht sehr viel schlechter als die beiden Verfahren, die auf Validitätsdaten beruhen. In der Regel scheint die Trennschärfe gerade für den Zweck am wenigsten geeignet zu sein, für den sie am häufigsten verwendet wird: zur Identifizierung einiger weniger Items, welche die Qualität des Tests beeinträchtigen.

Die Optimierung von Cronbachs α ist allenfalls bei einem eindimensionalen Itempool als Selektionsstrategie zu empfehlen. Bei einem mehrdimensionalen Itempool resultieren dagegen Skalen, die zwar eine sehr hohe Reliabilität, aber nur eine sehr unbefriedigende Validität erreichen. Je höher der Anteil der (systematischen) Fehlervarianz an der gemeinsamen Varianz der Items ist und je mehr er sich auf wenige Faktoren konzentriert desto weniger valide sind die resultierenden Skalen.

5.3.2 Reliabilität

Wenn man von der Zufallsauswahl absieht, waren die Reliabilitätskoeffizienten der Verfahren in Studie 3 umso höher je geringer die Validität der ausgewählten Tests war. Das MAXVAL-

[13] Bei der Entfernung von weiteren Items vermindert sich die Trennschärfe häufig jedoch kaum, wie man an dem Plateau zahlreicher Validitätscharakteristiken erkennen kann.

Verfahren erreichte mitunter sogar eine geringere Reliabilität als die Zufallsauswahl. Im Vergleich zu den anderen Verfahren erreichte das MAXVAL-Verfahren eine besonders geringe Reliabilität, wenn der Anteil der gemeinsamen Fehlervarianz groß oder die Kommunalität des Kriteriums und der Stichprobenumfang gering waren. Wenn eine höhere Anzahl von latenten Dimensionen, so wie in Studie 3b und 3c, nicht mit einem höheren Anteil gemeinsamer Fehlervarianz verbunden ist, so hat dies kaum Konsequenzen für die Reliabilität.

Bei Selektion anhand der Itemvalidität resultieren bei umfangreichen Itempools meist Skalen, die fast so reliabel sind wie bei Selektion anhand der Trennschärfe oder bei Optimierung von Cronbachs α. Auch bei weniger umfangreichen Itempools ist die Reliabilität bei Selektion anhand der Itemvalidität nicht wesentlich geringer, wenn der Anteil der gemeinsamen Fehlervarianz nicht zu groß ist und sich auf viele Faktoren verteilt. Dass die Reliabilität bei Selektion anhand der Itemvalidität bei einer höheren Anzahl latenter Dimensionen vergleichsweise höher ist, könnte daran liegen, dass eine gravierende (unsystematische) Kumulation der gemeinsamen Fehlervarianz unwahrscheinlicher ist, wenn sich diese auf viele Faktoren verteilt.

In Studie 2 gab es keine gemeinsamen Fehlerfaktoren. Die Items enthielten im Mittel genauso viel spezifische wie wahre Varianz, wobei die spezifische Varianz – genau wie in Studie 3 – als Fehlervarianz im Sinne der klassischen Testtheorie interpretiert wurde. Nur in Studie 2 gab es keinen wesentlichen Unterschied zwischen den Ergebnissen zur Reliabilität und denen zur Validität. Nur bei einem eindimensionalen Itempool steht man demnach nicht vor dem Dilemma, ob man sich auf die Sicherung der Validität oder der Reliabilität eines Tests konzentriert (vgl. Kapitel 1.3, S. 22).

In Studie 1 wurde die Reliabilität der ausgewählten Tests nicht ermittelt, da keine spezifischen Varianzquellen vorhanden waren. Wenn man Korrelationen zwischen den Messfehlern der Items ausschließt, so impliziert die Abwesenheit spezifischer Varianzanteile, dass die Reliabilität aller Items gleich eins ist. Demzufolge wäre auch die Reliabilität aller ausgewählten Subtests gleich eins. Da die Annahme einer perfekten Itemreliabilität ziemlich unrealistisch ist, könnte man geneigt sein, die Ergebnisse von Studie 1 als irrelevant zu betrachten. Allerdings ist auch die Annahme der Unkorreliertheit der Fehler nicht unbedingt sehr realitätsnah. Wenn man Korrelationen zwischen den Messfehlern im Sinne der klassischen Testtheorie zulässt, so kann die Reliabilität der Items und der ausgewählten Skalen auch bei Fehlen spezifischer

Varianzanteile kleiner als eins sein. Fehlervarianz im Sinne der klassischen Testtheorie kann in Studie 1 also nur Teil der gemeinsamen Fehlervarianz[14] der Items sein. Da in Studie 1 die gemeinsame Fehlervarianz nicht explizit in wahre und Messfehlervarianz aufgeteilt wurde, konnte die Reliabilität der Skalen nicht berechnet werden.

5.3.3 Suppression

In Kapitel 3.2 wurde ein neues Kriterium für Suppression im Rahmen der Testtheorie vorgestellt ([3.2-4] auf S. 44). Es zeigt sich, dass dieses Suppressionskonzept gute Dienste bei der Interpretation der Ergebnisse der Simulationsstudie liefert. So scheint ein gutes Abschneiden des MAXVAL-Verfahrens, wie erwartet, auf Fehlersuppression zu beruhen, während die teilweise recht bescheidene Validität bei Selektion anhand der Trennschärfe und Optimierung der internen Konsistenz auf Fehlerredundanz zurückführbar sein dürfte. Die Selektion anhand der Itemvalidität begünstigt dagegen ebenso wenig wie die zufällige Auswahl das Vorkommen von Fehlersuppression oder Fehlerredundanz.

Je höher der Anteil der gemeinsamen Fehlervarianz desto stärker ist die Fehlerredundanz bei Selektion anhand der Trennschärfe und der Optimierung der internen Konsistenz. Dies geht mit einer geringen Validität der beiden Verfahren einher. Bei konstantem Anteil der gemeinsamen Fehlervarianz geht eine höhere Anzahl gemeinsamer Faktoren nicht mit einer höheren Fehlerredundanz einher. Bei Optimierung von Cronbachs α scheint die Fehlerredundanz sogar mit der Anzahl der latenten Dimensionen abzunehmen.

Die Fehlersuppression bei Selektion mithilfe des MAXVAL-Verfahrens scheint dagegen nicht mit dem Varianzanteil der gemeinsamen Fehlervarianz, sondern mit der Anzahl der gemeinsamen Fehlerfaktoren zuzunehmen. Das schlechtere Abschneiden des MAXVAL-Verfahrens in Studie 3b in Vergleich zu Studie 3a scheint also eher auf die mangelnde Kumulation von valider Varianz zurückzuführen zu sein als auf eine geringere Wirksamkeit von Suppressionseffekten. Selbst bei einem eindimensionalen Itempool versucht das MAXVAL-Verfahren von vermeintlichen Suppressionseffekten zu profitieren, obwohl Suppressionseffekte hier gar nicht vorkommen können. Dies scheint die Ursache für die unbefriedigende Validität des MAXVAL-Verfahren in Studie 2 zu sein.

[14] In Kapitel 4.2 wurde jegliche kriteriumsirrelevante Varianz als gemeinsame Fehlervarianz definiert.

Wie soeben angedeutet, zeigt das in Kapitel 3.2 vorgeschlagene Suppressionsmaß ([3.2-4] auf S. 44) auch bei einem eindimensionalen Itempool beim MAXVAL-Verfahren tendenziell Fehlersuppression an, während es bei Selektion anhand der Trennschärfe und der Optimierung von Cronbachs α mitunter Fehlerredundanz bescheinigt. Tatsächlich kann aber bei Eindimensionalität *in der Population* weder Fehlerredundanz noch Fehlersuppression vorliegen. Die irreführenden Werte sind das Ergebnis von Stichprobenfehlern bei der Schätzung der Kovarianzmatrix. Die Stichprobenfehler dürften jedoch nur dann in systematische Weise zu falschen Schlussfolgerungen über die Verhältnisse in der Population führen, wenn das Suppressionsmaß – so wie in Studie 2 – in derselben Stichprobe berechnet wird, die auch zur Itemselektion herangezogen wurde[15]. Daher sollte das hier vorgestellte Suppressionskriterium in der Praxis nur in einer neuen Validierungsstichprobe zur Anwendung kommen.

5.3.4 Ökologische Validität der Ergebnisse

Die Aussagekraft der Ergebnisse von Simulationsstudien hängt entscheidend davon ab, inwieweit es gelingt, Szenarien zu modellieren, die möglichst realitätsnah sind. Die Realitätsnähe lässt sich wiederum danach beurteilen, ob das gewählte Modell überhaupt für den Gegenstandsbereich angemessen ist, und ob die Parameter innerhalb des Modells so gesetzt wurden, dass die realen Verhältnisse möglichst gut repräsentiert werden.

Angemessenheit des Modells

Als Modell wurde in der vorliegenden Simulationsstudie auf die Faktorenanalyse zurückgegriffen. Man kann die Faktorenanalyse als mehrdimensionale Erweiterung von Modellen der klassischen Testtheorie verstehen (vgl. Kapitel 1.1.1, S. 11). Die Faktorenanalyse wird auch häufig als Methode zur Konstruktion und Analyse von psychologischen Testverfahren eingesetzt. Es ist jedoch fraglich, ob das Modell der Faktorenanalyse für die Analyse und Beschreibung der Daten psychologischer Tests auf Itemebene wirklich geeignet ist. Die implizite Annahme intervallskalierter Daten ist dabei noch das geringere Problem, da man bei Items mit geordneten Kategorien durch die Einführung von Zählvariablen sogar Absolutskalenniveau unterstellen kann (Krauth, 1995). Schwerwiegender ist die Tatsache, dass die Itemscores in der Regel diskret sind. Formal betrachtet kann die Faktorenanalyse auch diskrete Verteilungen der manifesten Itemscores abbilden. Dazu sind jedoch sehr spezielle Annahmen über die gemeinsame Verteilung der Faktorwerte (auf den gemeinsamen und spezifischen Faktoren) und

[15] In Studie 3 wurde das Suppressionsmaß jeweils anhand der Kovarianzmatrix in der Population berechnet. In beiden Fällen wurde dabei die Kommunalität des Kriteriums als bekannt vorausgesetzt.

der Messfehler erforderlich. Es ist ein zentrales Kennzeichen klassischer testtheoretischer Modelle, dass das Itemformat der Items nicht in angemessener Weise berücksichtigt wird. Bei Modellen der Item-Response Theorie wird das Itemformat dagegen immer bei der Modellbildung berücksichtigt. Mehrdimensionale, probabilistische Modelle befinden sich jedoch erst im Stadium der Entwicklung (vgl. Rost, 1996, 1999). Daher wurde trotz der methodischen Bedenken auf das klassische Modell der Faktorenanalyse zurückgegriffen. Dementsprechend ist die Verteilung der Itemscores kontinuierlich. Ob sich die Ergebnisse der vorliegenden Simulationsstudie auf Items mit wenigen Abstufungen übertragen lassen, sollte daher in weiteren Simulationsstudien überprüft werden.

Bei klassischer Auswertung von diskreten Testitems, ergeben sich im Rahmen der Faktorenanalyse zwangsläufig Schwierigkeitsfaktoren[16] (Lienert & Raatz, 1994), wenn die Randhäufigkeiten der Items unterschiedlich sind. Es sind also in der Regel mehr Dimensionen zur Beschreibung der Daten notwendig als man aufgrund der Anzahl der latenten, psychologischen Dimensionen erwarten würde. Daher dürfte τ-Kongenerität (Studie 2) bei diskreten Testitems mit wenigen Abstufungen kaum vorkommen. Selbst die Items eines Rasch-homogenen Itempools sind in der Regel nicht eindimensional im Sinne der Faktorenanalyse.

Angemessenheit der Parametersetzungen

In der vorliegenden Simulationsstudie wurde versucht, den Einfluss von möglichst vielen Faktoren auf Effektivität der verschiedenen Selektionsverfahren bei der Sicherung der Validität und Reliabilität zu untersuchen. Dabei wurde angestrebt, den Wertebereich der Parameter so zu variieren, dass er Werte, wie sie in empirischen Anwendungen vorkommen, nach Möglichkeit einschließt. Daher lassen sich recht differenzierte Aussagen darüber machen, unter welchen Bedingungen, welches Verfahren besonders angemessen erscheint (siehe Kapitel 5.3.1, S. 139 bis 5.3.3, S. 143). Bei einer feineren Abstufung der Faktoren wären zwar noch präzisere Empfehlungen für die Testkonstruktion möglich gewesen. Dies hätte jedoch den Rahmen der vorliegenden Arbeit gesprengt.

Trotz aller Sorgfalt bei der Planung der Szenarien für die Simulationsstudien, bleibt zu bedenken, dass die Ergebnisse von Simulationsstudie letztlich immer auf Idealisierungen und

[16] Die von den Schwierigkeitsfaktoren aufgeklärte Varianz hängt freilich von dem Ausmaß der Unterschiede in den Randhäufigkeiten ab, so dass in der Empirie nicht unbedingt markante Schwierigkeitsfaktoren zu beobachten sind.

einer Vergröberung der realen Verhältnisse beruhen. Dies soll im Folgenden an einigen Punkten demonstriert werden:

Der Algorithmus zur Bestimmung der Ladungen auf den gemeinsamen Fehlerfaktoren führt dazu, dass positive Ladungen im Mittel in etwa genau so häufig vorkommen wie negative Ladungen. In der Praxis ist jedoch damit zu rechnen, dass es eine Tendenz zu positiven oder negativen Ladungen gibt. Wenn man beispielsweise davon ausgeht, dass die Tendenz zu sozial erwünschtem Antwortverhalten ein möglicher gemeinsamer Fehlerfaktor ist, so würde man bei einer sozial erwünschten Eigenschaft überwiegend positive Ladungen auf diesem Faktor erwarten. Bei einem Test zur Erfassung einer sozial unerwünschten Eigenschaft ist dementsprechend mit vorwiegend negativen Ladungen zu rechnen. Unter diesen Umständen dürfte es bei Selektion anhand der Itemvalidität, der Itemtrennschärfe sowie bei der Optimierung von Cronbachs α zu einer deutlichen und *systematischen* Kumulation von kriteriumsirrelevanter Varianz kommen. Auch bei zufälliger Auswahl der Items dürfte es, anders als in den vorliegenden Simulationsstudien, eine Tendenz zur Fehlerredundanz geben. Das MAXVAL-Verfahren würde unter diesen Umständen über die Suppression der gemeinsamen Fehlervarianz vermutlich besonders hohe Validitätskoeffizienten erreichen.

Die Ladungen auf dem validen Faktor sowie auf den gemeinsamen Fehlerfaktoren variieren in der Simulationsstudie von Item zu Item. Die Ladung auf dem spezifischen Faktor war dagegen bei allen Items innerhalb einer Simulationsstudie konstant. Damit sollte die Häufigkeit von extrem hohen Itemvaliditäten reduziert werden. Wenn man die Ladung auf dem spezifischen Faktor als Messfehler im Sinne der klassischen Testtheorie interpretiert, dann implizieren homogene Ladungen die Varianzhomogenität der Fehler[17]. Auch diese Annahme dürfte in der Praxis nicht immer erfüllt sein. Es ist schwer abzuschätzen, welche Konsequenz die Heterogenität der Fehlervarianz auf den Vergleich der verschiedenen Selektionsverfahren hat. In Studie 2 hat die Annahme der Varianzhomogenität jedoch mit Sicherheit das gute Abschneiden der Selektion anhand der Trennschärfe und der Itemvalidität begünstigt, da es hier nur bei heterogenen Fehlervarianzen Sinn machen kann, ein weniger valides und trennscharfes Item vorzuziehen (z.B. wenn die Varianz des valideren Items so gering ist, dass das Item keinen nennenswerten Einfluss auf den Skalensummenwert hat).

[17] Gemeint ist hier die Varianzhomogenität beim Vergleich der Items. Sie dürfte besonders bei Boden- und Deckeneffekten nicht gegeben sein. Die Annahme der Varianzhomogenität beim Vergleich der Messfehler verschiedener Personen, wie sie für Zwecke der Einzelfalldiagnostik bei der Ermittlung von Konfidenzintervallen gemacht wird, ist hier nicht gemeint.

Die Varianzaufklärung der gemeinsamen Fehlerfaktoren hat sich in der vorliegenden Simulationsstudie nicht (systematisch) unterschieden. Daher ist der Eigenwerteverlauf der Korrelationsmatrix der Items (in der Population!) treppenförmig mit drei Stufen (valider Faktor – gemeinsame Fehlerfaktoren – spezifische Faktoren, vgl. Basilevsky, 1994). Dies dürfte so in der Realität der Testkonstruktion wohl eher selten vorkommen. Die Ergebnisse der Studien 3b und 3c lassen jedoch vermuten, dass eine heterogene Verteilung der Varianz der Fehlerfaktoren dem MAXVAL-Verfahren zugute käme. Bei extremer Heterogenität würde sich nämlich die gesamte gemeinsame Fehlervarianz auf einem einzigen Faktor konzentrieren. Unter diesen Umständen hat das MAXVAL-Verfahren etwas besser abgeschnitten, als bei einer gleichmäßigen Verteilung der gemeinsamen Fehlervarianz über alle Faktoren.

In der vorliegenden Simulationsstudie wurden die Items immer so gepolt, dass sie in der Stichprobe eine positive Validität haben[18]. Bei der üblichen Testkonstruktion anhand von Itemanalysen stehen in der Regel jedoch gar keine Daten über die Validität der Items zur Verfügung. Daher dürfte es häufiger auch zu negativen Itemvaliditäten kommen. Dies wirkt sich vermutlich negativ auf die Validität von Tests aus, bei denen die Items anhand der Trennschärfe oder Cronbachs α ausgewählt werden.

[18] Da die Polung der Items nicht von der Validität in der Population, sondern von der Validität in der Stichprobe abhing, war – v.a. bei geringem Stichprobenumfang – auch für die Validität des Skalensummenwerts aller Items des Itempools eine Regression zur Mitte zu beobachten.

6 Empirische Studie

Der im vorangegangen Kapitel angestellte Vergleich der untersuchten Itemselektionsalgorithmen anhand von simulierten Daten lieferte eine Reihe interessanter Ergebnisse. Da die Repräsentativität simulierter Datensätze für die in der Empirie vorkommenden Daten auch bei größter Sorgfalt bei der Setzung der Parameter nicht als gegeben vorausgesetzt werden kann, sollen die verschiedenen Selektionsalgorithmen auch anhand empirischer Datensätzen verglichen werden. Dieses Vorgehen soll Hinweise über die externe Validität der beobachten Ergebnisse liefern. Da bei empirischen Datensätzen die Populationsparameter unbekannt sind, lassen sich die im vorangegangen Kapitel untersuchten Hypothesen anhand der empirischen Daten jedoch teilweise gar nicht untersuchen. Die andere Natur der empirischen Daten erforderte zum Teil ein modifiziertes methodisches Vorgehen.

6.1 Methode

6.1.1 Stichproben

Für die empirische Studie wurde auf den Datensatz von Amelang und Borkenau (1981) zurückgegriffen. Die Daten stammen von insgesamt 424 Personen. Für 344 Personen lagen Fremdratings von je drei ihnen gut bekannten Personen auf 32 Merkmalsdimensionen vor. Bei 321 Personen waren auch alle entsprechenden Selbstratings vorhanden. Die Fremdbeurteilungen wurden genau wie die Selbstratings in siebenfacher Abstufung erhoben. Außerdem wurde den Testpersonen eine Vielzahl von Persönlichkeitstests vorgegeben. Insgesamt dauerte die Bearbeitung des Testmaterials etwa 5 Stunden und fand an zwei verschiedenen Tagen in den Räumen des psychologischen Instituts der Universität Heidelbergs statt. Die Testpersonen erhielten 70,- DM für die vollständige Bearbeitung des Fragebogens; die Fremdbeurteiler erhielten 10,- DM. Eine detaillierte Beschreibung von Stichprobe und Design findet man bei Borkenau (1981).

Für die Analysen wurden neben den Selbst- und Fremdratings die Items der Gesamtform des Freiburger Persönlichkeitsinventars (FPI, Fahrenberg, Selg & Hampel, 1973) verwendet. Lediglich die Items aus der Skala „Offenheit“ gingen nicht in die Analysen mit ein, da sie von Amelang und Borkenau (1981) nicht verwendet wurden und somit keine Information über deren Validität vorlag. Für die Auswahl des FPI sprach vor allem der Umstand, dass bis auf die Skalen „Nervosität“ und „Offenheit“ für jede der acht vom FPI erfassten Persönlichkeitsdimension Fremdratings vorlagen, die direkt den Labels entsprachen, die im Handbuch des FPI den einzelnen Skalen zugeordnet werden. Da die Skala "Nervosität" vor allem das Ausmaß

psychosomatischer Beeinträchtigungen erfasst, wurde ihr die Beurteilungsdimension "Körperliche Labilität" zugewiesen. Für eine Verwendung des FPI sprach außerdem der relativ breite Persönlichkeitsbereich, der mit Hilfe des FPI erfasst wird, sowie die erhebliche Variabilität der psychometrischen Qualität der Items in der Originalversion des FPI[19]. Variabilität in der psychometrischen Qualität des Itempools ist die Voraussetzung dafür, dass sich überhaupt Unterschiede in der Güte einzelner Selektionsstrategien ergeben können. Vor Bearbeitung des FPI hatten die Testpersonen bereits 405 Items aus anderen Skalen bearbeitet.

6.1.2 Analysen

Alle in Kapitel 4.1 (S. 48) erwähnten Selektionsmethoden bis auf die vollständige Permutation wurden auf die empirischen Daten angewendet. Der Vergleich der verschiedenen Selektionsalgorithmen erfolgte, indem jedes der Selektionskriterien auf den jeweiligen Itempool angewendet wurde. Dazu wurden für jede Testlänge die Items bestimmt, die nach Maßgabe des jeweiligen Selektionsalgorithmus ausgewählt wurden und anschließend die externe Validität der selegierten Subtests verglichen. Als Validitätskriterium wurde dabei die korrelative Übereinstimmung mit dem Mittelwert der drei Fremdbeurteilungen herangezogen. Schätzungen der Reliabilität des Mittelwerts der Fremdratings schwankten zwischen .37 und .72. Da einige der Selektionsalgorithmen bereits auf Validitätsdaten basieren, wurde die Personenstichprobe zuvor per Zufall in eine Analysestichprobe und eine Validierungsstichprobe aufgeteilt. Die Itemselektion erfolgte aufgrund der Daten der Analysestichprobe, während die Bestimmung der Validität auf den Daten der Validierungsstichprobe beruhte. Da aufgrund der Regression zur Mitte nur bei einer hinreichend großen Personenstichprobe Verbesserungen durch multivariate Selektionsmethoden zu erwarten sind, wurden zwei Drittel der Personen der Analysestichprobe zugewiesen. Der hierdurch bedingte Verlust an Präzision in der Validierungsstichprobe betrifft dagegen alle Selektionsmethoden gleichermaßen, so dass kein systematischer Effekt auf den Vergleich der Selektionsmethoden zu erwarten ist.

Als Itempool wurden zum einen die 32 Selbstratings und zum anderen die Items der acht Skalen des FPI verwendet. Die Selbstratings wurden vor der Itemselektion jeweils so gepolt, dass sie in der Analysestichprobe eine positive Korrelation mit demjenigen Fremdrating haben, das gerade als Validitätskriterium diente. Die Ergebnisse der Selbstratings und der FPI-Skalen werden

[19] Da die Konstrukteure des FPI zwei Parallelformen anbieten wollten, wurden von 257 Items, die ursprünglich zur Testkonstruktion zur Verfügung standen, nur 47 ausgesondert, obwohl die psychometrische Qualität einiger Items zu wünschen übrig ließ. Daher dürfte sich die Gesamtform des FPI nicht allzu sehr von Itempools unterscheiden, die üblicherweise zur Testkonstruktion verwendet werden.

getrennt präsentiert, da die 32 Selbstratings einen sehr heterogenen, mehrdimensionalen Itempool darstellen, während die Items der acht Skalen des FPI aufgrund ihrer faktoranalytischen Herkunft sehr viel homogener sein dürften.

Wie bereits erwähnt, wurde bei den acht FPI-Skalen die Korrelation mit demjenigen globalen Fremdrating, welches dem Label der FPI-Skala entspricht, als Validitätskennwert verwendet. Die verschiedenen Selektionskriterien konnten demnach an acht verschiedenen Itemgrundmengen mit jeweils unterschiedlichem Validitätskriterium verglichen werden. Dennoch sind die Ergebnisse nicht in strengem Sinne unabhängig, da sie jeweils anhand derselben Personenstichprobe gewonnen wurden.

Die 32 Selbstratings wurden sogar mehrfach als Itempool verwendet, da die Korrelation mit jedem der 32 Fremdratings als Validitätskriterium verwendet wurde. Dies ermöglicht zwar, den Vergleich der verschiedenen Selektionskriterien auf eine breitere empirische Basis zu stellen; es bleibt jedoch zu bedenken, dass sich die Kovarianzmatrizen, die zur Itemselektion und der Berechnung der Validität herangezogen werden, (bis auf etwaige Umpolungen der Items) jeweils nur in einer Zeile (bzw. Spalte) voneinander unterscheiden. Dennoch sind die diesbezüglichen Ergebnisse keineswegs vollständig redundant, da der Vergleich der einzelnen Selektionsalgorithmen dennoch zu unterschiedlichen Ergebnissen führen kann. Insbesondere kann man nicht davon ausgehen, dass Suppressionsbeziehungen zwischen den Items bei der Vorhersage der verschiedenen Fremdratings von gleichbleibender Bedeutung sind, da Suppression immer nur hinsichtlich eines bestimmten Kriteriums definiert ist (vgl. [3.2-6] auf S. 44). Unter methodischen Gesichtspunkten wäre es natürlich wünschenswert, wenn man die verschiedenen Selektionsalgorithmen an einer Vielzahl von empirischen Datensätzen mit jeweils unterschiedlichen Personen- und Itemstichproben untersuchen könnte. Ein solches Vorhaben ist jedoch praktisch kaum realisierbar, da der Bedarf an Versuchsperson exorbitant wäre.

6.2 Ergebnisse

Die Ergebnisse bei vollständiger Permutation und die Optimierung des MAXVAL-Verfahrens lieferten nahezu dieselben Resultate wie das einfache MAXVAL-Verfahren von Ulrich (1985). Daher wird auf eine gesonderte Darstellung verzichtet.

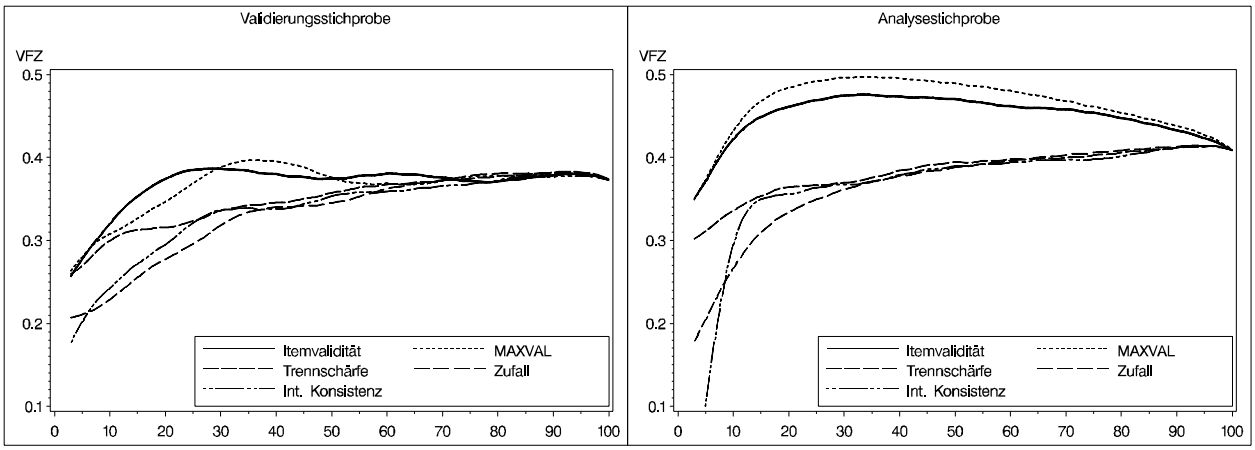
6.2.1 FPI-Skalen

Die beiden Verfahren, die auf Validitätsdaten basieren, erreichen in der Analysestichprobe deutlich höhere Validitäten als die Selektion anhand der Trennschärfe und die Optimierung der internen Konsistenz (vgl. Abbildung 61 auf S. 151). Nur bei sehr kurzen Skalen erreichen

zufällig ausgewählte Skalen eine geringere Validität als bei Selektion anhand der Trennschärfe oder bei Optimierung der internen Konsistenz.

In der Validierungsstichprobe erreichen die Verfahren, die auf Validitätsdaten basieren, dagegen nur bei kurzen Skalen höhere Validitäten. Bei Optimierung der internen Konsistenz werden Validitäten in derselben Größenordnung erreicht, wie bei zufälliger Auswahl. Außer bei sehr kurzen Skalen sind auch bei Selektion anhand der Trennschärfe kaum höhere Validitäten zu beobachten als bei zufälliger Auswahl. Wenn mehr als die Hälfte der Items in den Test aufgenommen wurde, dann unterscheidet sich keines der Verfahren mehr deutlich von der Zufallsauswahl.

Abbildung 61**:** Validität der verschiedenen Selektionsverfahren bei der Auswahl von Items des FPI



Auf der Abszisse ist dargestellt, wie viel Prozent der Items in den Test aufgenommen wurden. Da die Skalen des FPI unterschiedlich lang sind, erfolgte die Anpassung der Regressionslinie mithilfe der kubischen Spline-Methode von Reinsch (1967).

Bei der graphischen Darstellung der empirischen Ergebnisse wurden bisher immer Mittelwerte von mehreren Tests verglichen, die in unterschiedlichen Itemgrundmengen (Skalen) zu beobachten waren. Innerhalb eines einzelnen Itempools sind jedoch zum Teil erhebliche Abweichungen vom allgemeinen Trend zu beobachten[20]. Um dies zu veranschaulichen wurde in Abbildung 62 (S. 152) und Abbildung 63 (S. 153) dargestellt, welche Validität bei Anwendung der verschiedenen Selektionsverfahren in den einzelnen Skalen des FPI in der Analyse- und Validierungsstichprobe zu beobachten war.

[20] Daher waren die in Abbildung 61 (S. 151) dargestellten Unterschiede zwischen den Selektionsverfahren bei der geringen Zahl der verglichenen Itemgrundmengen auch nicht statistisch signifikant.

Abbildung 62: Validität (Fisher-Z standardisiert) in der Analysestichprobe bei Selektion innerhalb der einzelnen FPI- Skalen

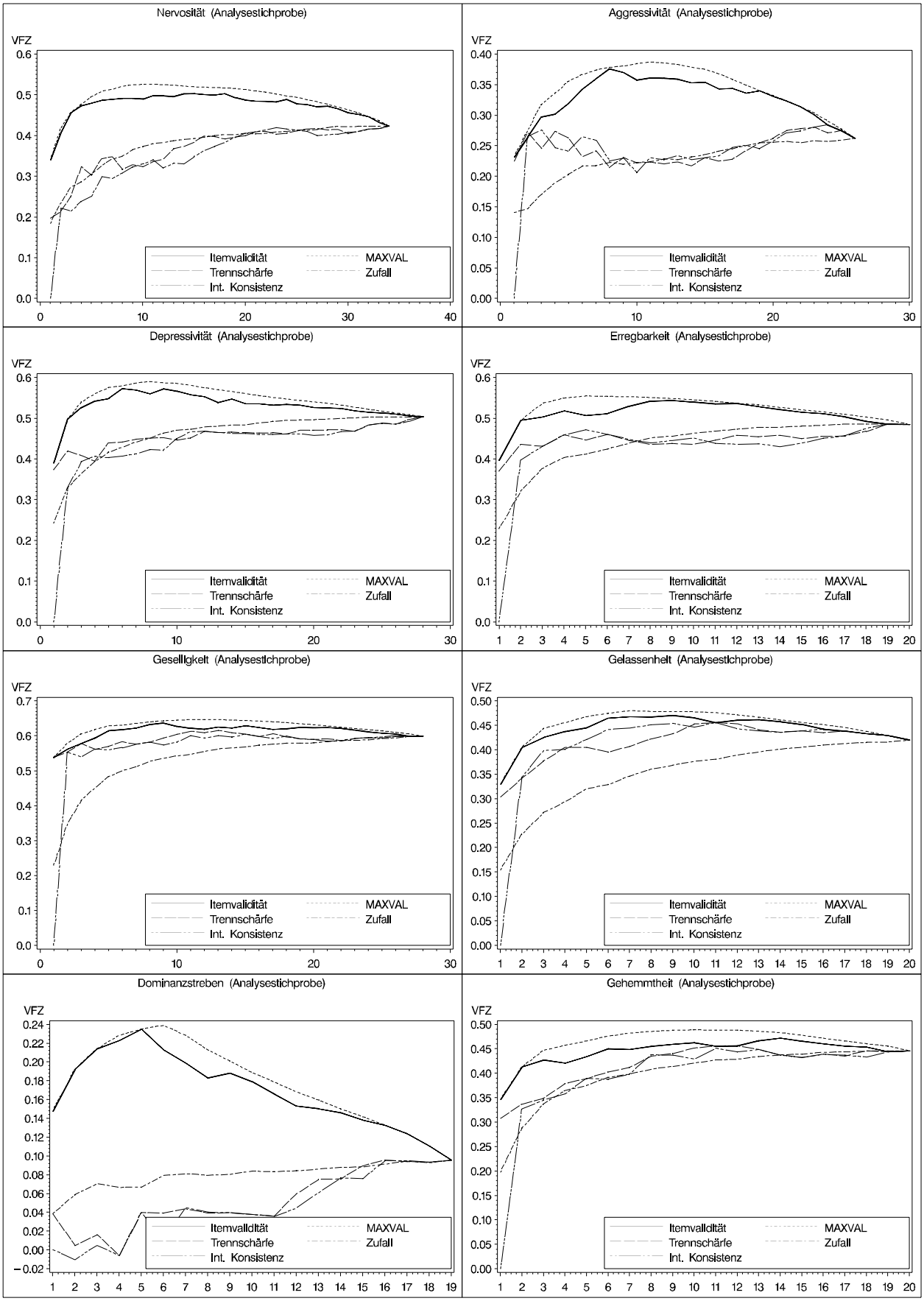
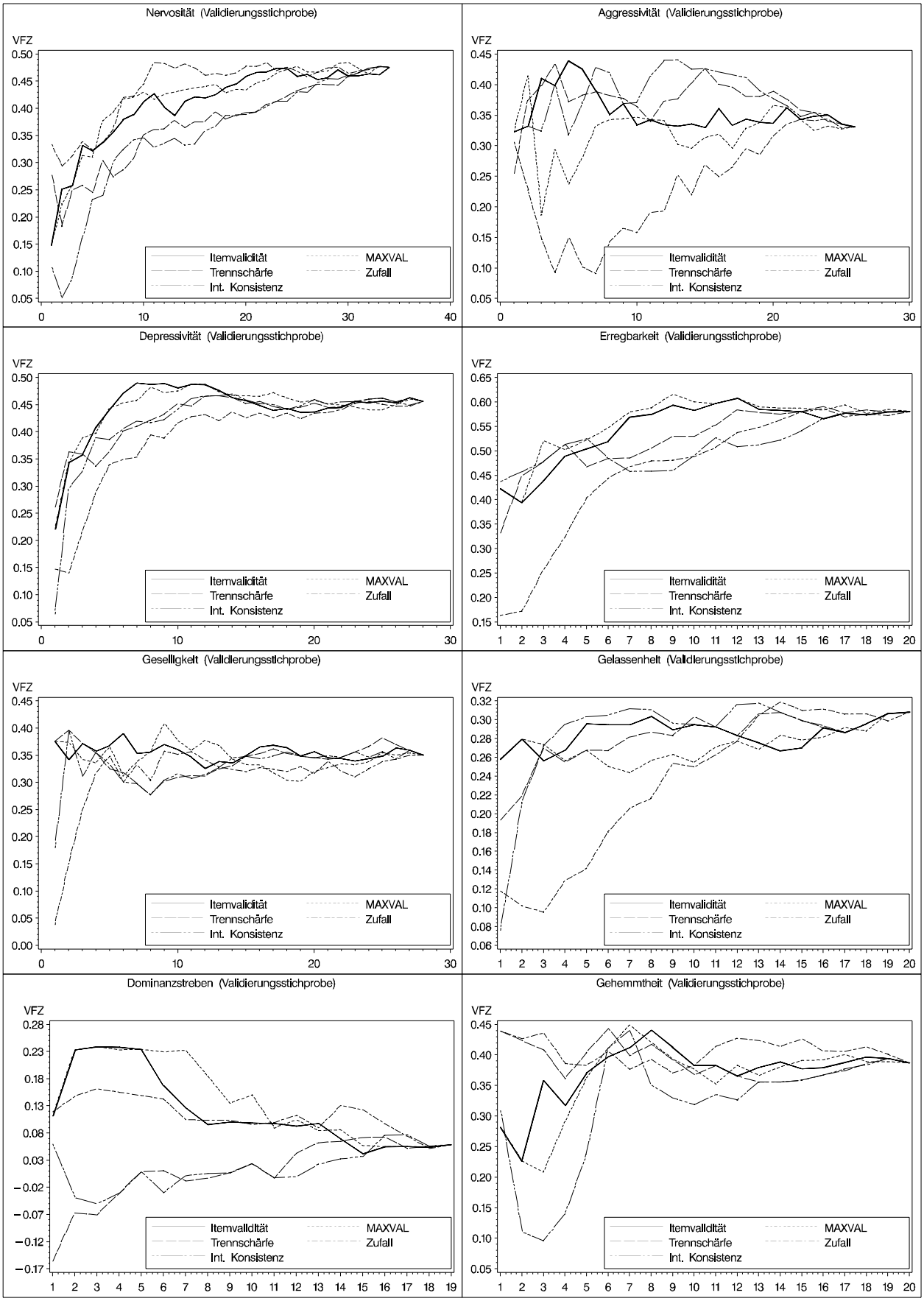


Abbildung 63: Validität (Fisher-Z standardisiert) in der Validierungsstichprobe bei Selektion innerhalb der einzelnen FPI- Skalen

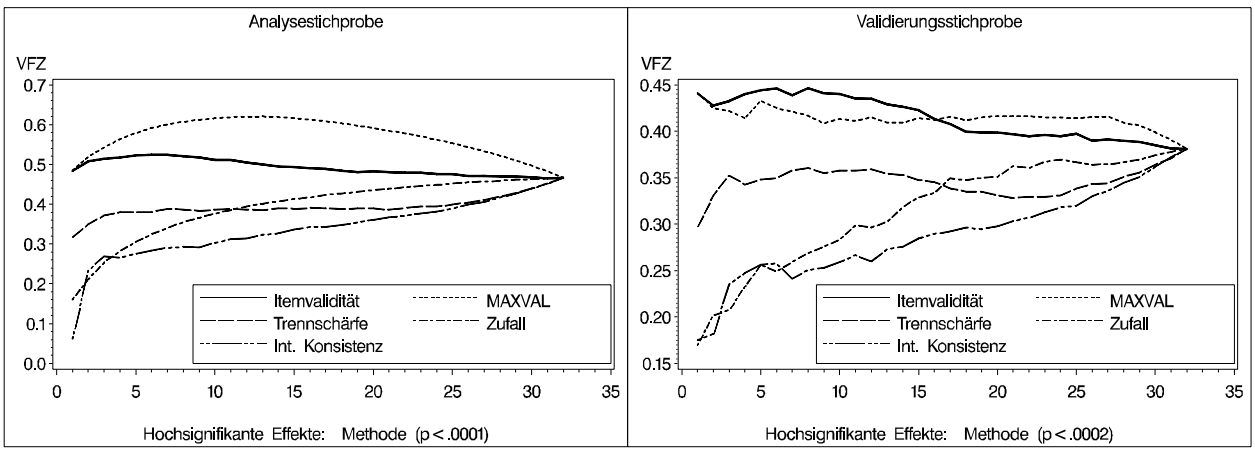


Bei der Betrachtung der Graphiken sollte man berücksichtigen, dass die Korrelationen sowohl in der Validierungsstichprobe als auch in der Analysestichprobe mit einem erheblichen Stichprobenfehler geschätzt wurden. Das 95%-Konfidenzintervall der Fisher-Z Korrelationen umfasst bei einem Stichprobenumfang von 107 (Validierungsstichprobe) ein Intervall mit der Breite .38 und bei einem Stichprobenumfang von 214 (Analysestichprobe) immerhin noch ein Intervall mit der Breite .27 (Bortz, 1999). Der Verlauf der Validitätscharakteristiken ist in der Analysestichprobe wesentlich gleichmäßiger als in der Validierungsstichprobe. Dies ist zumindest zum Teil auf den größeren Stichprobenumfang zurückzuführen.

6.2.2 Selbstratings

Bei Anwendung der verschiedenen Selektionsalgorithmen auf die Auswahl von globalen Selbstratings fällt auf, dass die Items mit der höchsten Validität in der Analysestichprobe in der Validierungsstichprobe eine höhere Validität haben als die Summe aus allen 32 Selbstratings (vgl. Abbildung 64). Ansonsten konnten die Ergebnisse aus der Simulationsstudie in mehrfacher Hinsicht bestätigt werden. Auch anhand der empirischen Daten zeigte sich, dass bei Selektion anhand der Trennschärfe vergleichsweise hohe Validitäten in der Validierungsstichprobe resultieren, wenn nur einige wenige Items ausgewählt werden. Wenn dagegen ein großer Anteil der Items in den Test aufgenommen wird, dann führt die Selektion anhand der Trennschärfe nicht zu besseren Validitäten als bei einer zufälligen Auswahl der Items. Bei den Selbstratings war die zufällige Auswahl sogar etwas besser als die Selektion anhand der Trennschärfe, wenn mehr als die Hälfte der Items aufgenommen wurde. Eine Optimierung der internen Konsistenz führte zu noch schlechteren Validitäten als bei zufälliger Auswahl.

Abbildung 64: Validität der verschiedenen Selektionsverfahren bei Auswahl von globalen Selbstratings



6.3 Interpretation

Bei der Interpretation der Ergebnisse der in diesem Kapitel dargestellten empirischen Studie ist Vorsicht angebracht, da die Itemselektion jeweils anhand von Daten aus einer einzigen Personenstichprobe vorgenommen wurde. Bei den Selbstratings war sogar der Itempool bis auf etwaige Umpolungen der einzelnen Selbstratings identisch. Folglich kann man nicht davon ausgehen, dass die Vergleiche der verschiedenen Selektionsverfahren innerhalb der insgesamt 40 untersuchten Itemgrundmengen unabhängig voneinander sind. Demnach sind auch die statistischen Voraussetzungen für einen Vergleich der verschiedenen Selektionsmethoden verletzt. Die Ergebnisse der empirischen Studie haben daher eher deskriptiven Charakter. Sie zeigen, dass viele der im Methodenteil postulierten und in der Simulationsstudie gefundenen Effekte sich auch bei empirischen Daten beobachten lassen.

Aus den Ergebnissen der Simulationsstudie lässt sich entnehmen, welche Ergebnisse in der empirischen Studie zu erwarten waren. So ist bei einem Stichprobenumfang von etwa 200 Personen und einer Reliabilität der Fremdratings von etwa .55 durchaus noch eine substantielle Regression zur Mitte zu erwarten, die insbesondere beim MAXVAL-Verfahren zu Validitätseinbußen führen dürfte. Bei den Selbstratings kann man davon ausgehen, dass der Itempool eher heterogen ist. Der Anteil der gemeinsamen Fehlervarianz dürfte deutlich höher sein als der Anteil der validen Varianz. Die FPI-Skalen dürften zwar deutlich homogener sein, da sie faktorenanalytisch konstruiert wurden, aber man kann dennoch davon ausgehen, dass der Itempool nicht eindimensional ist, da die Konstrukteure des FPI bei der Testentwicklung relativ hohe Ladungen der Items auf fremden Faktoren tolerierten. Von daher war zu erwarten, dass die Optimierung von Cronbachs α und die Selektion anhand der Trennschärfe keine besonders guten Ergebnisse liefern.

Tatsächlich konnten viele Ergebnisse der Simulationsstudie sowie der eingangs vorgestellten analytischen Arbeiten anhand der empirischen Testdaten bestätigt werden. So war die Selektion anhand der Trennschärfe bei den Selbstratings tendenziell sogar schlechter als die zufällige Auswahl, wenn nur ein kleiner Teil der Items aus dem Test entfernt wurde. Der Validität der Skalen aus globalen Selbstratings fiel sogar tendenziell mit zunehmender Testlänge, wenn die Selbstratings anhand ihrer Validität in der Analysestichprobe ausgewählt wurden. Das MAXVAL-Verfahren und die Selektion anhand der Itemvalidität erreichten in der Analysestichprobe meist deutlich höhere Werte als die anderen Verfahren. Die Regression zur Mitte war beim MAXVAL-Verfahren nur bei den Selbstratings deutlicher ausgeprägt als bei der Selektion anhand der Itemvalidität.

Bei den Ergebnissen des MAXVAL-Verfahrens innerhalb der FPI-Skalen zeigten sich ähnliche Ergebnisse wie in der Studie von Burisch (1997). Auch hier waren nur in den jeweiligen Analysestichproben deutlich höhere Validitäten erreicht worden als wenn nur ein Teil der zur Auswahl stehenden Items in den Test aufgenommen wurde. Die Ergebnisse in den Validierungsstichproben waren dagegen weniger spektakulär. Angesichts der in der vorliegenden Arbeit berichteten Ergebnisse muss auch bezweifelt werden, dass das MAXVAL-Verfahren für die Auswahl von sehr kurzen Skalen besser geeignet ist als die Selektion anhand der Itemvalidität, wenn, wie in der Studie von Burisch (1997), mit Stichproben von etwa 100 Personen gearbeitet wird. Tatsächlich erreichten die von Burisch vorgestellten Kurzskalen in dem Datensatz von Amelang und Borkenau (1981) im Durchschnitt nur eine Validität von .30 im Gegensatz zu .39 in der Studie von Burisch. Für die Orginalskalen des FPI ergab sich eine mittlere Validität von .39, während Burisch eine mittlere Validität von .37 ermittelte.

Das globale Selbstrating, das in der Analysestichprobe die höchste Validität hatte, erreichte auch in der Validierungsstichprobe meist eine sehr hohe Validität. In der Regel bezog sich dieses Selbstrating auf dieselbe Beurteilungsdimension wie das Fremdrating, das gerade als Validitätskriterium diente. Die Aufnahme weiterer Selbstratings führte nur in der Analysestichprobe, nicht jedoch in der Validierungsstichprobe, zu Skalen mit wesentlich höherer Validität. Dies unterstützt die Ergebnisse von Yousfi (1999), die ebenfalls nahe legen, dass es in der Regel nicht möglich ist, aus den mit den üblichen Persönlichkeitsfragebögen gewonnenen Selbstbeschreibungsdaten, Persönlichkeitsmaße zu entwickeln, die valider sind als globale Selbstratings (vgl. auch Burisch, 1984a, 1984b).

Insgesamt muss jedoch festgehalten werden, dass ein Vergleich der verschiedenen Selektionsmethoden anhand der hier verwendeten Datensätze ebenso wenig zu verlässlichen Ergebnissen führt, wie die Daten, die Burisch (1997) zur Verfügung standen. Dazu müsste eine Vielzahl von Itemgrundmengen an jeweils unterschiedlichen Personenstichproben untersucht werden. Da solch umfangreiche Datensätze kaum von einem einzigen Forscher erhoben werden können, lassen sich aussagekräftige Ergebnisse, die nicht nur deskriptiven Charakter haben, allenfalls durch Metaanalysen gewinnen. Bis solche Studien vorliegen, sollte man sich bei der Wahl der Selektionsmethode eher auf die Ergebnisse von Simulationsstudien stützen.

7 Diskussion

Ziel der vorliegenden Arbeit ist es, die übliche Praxis bei der Testkonstruktion kritisch zu hinterfragen und zu untersuchen, ob multivariate Methoden der Itemselektion zu Tests höherer Reliabilität und Validität führen. Dabei geht es weniger um die Frage, wie man Items generieren kann, die als Messinstrument für ein bestimmtes Merkmal geeignet sind, sondern vielmehr darum, wie sich aus einem bestehenden Itempool Skalen zusammenstellen lassen, die dieses Merkmal möglichst reliabel und valide erfassen. In der Regel versucht man dieses Ziel zu erreichen, indem man Skalen aus möglichst vielen, möglichst trennscharfen Items bildet. Betrachtet man die Testbesprechungen in den einschlägigen psychologischen Fachzeitschriften, so drängt sich der Eindruck auf, dass dieses Vorgehen der „state of the art" der Testkonstruktion ist.

Erstaunlicherweise lässt sich weder aus den Formeln der klassischen Testtheorie noch aus den Arbeiten zur Item-Response Theorie eine hinreichende Begründung für diese Praxis ableiten. Die vielzitierte Spearman-Brown Formel ([1.3-1] auf S. 21 und [1.3-2] auf S. 21) gilt nur für parallele Items bzw. essentiell τ-äquivalente Items, deren Fehler unkorreliert sind und jeweils dieselbe Varianz haben. Wie den Ausführungen in Kapitel 1.3 (ab S. 21) und den empirischen Ergebnissen in Kapitel 5 und 6 zu entnehmen ist, kann man nicht davon ausgehen, dass die Reliabilität und Validität eines Tests bei Verletzungen dieser Vorraussetzungen mit zunehmender Testlänge ansteigen[21]. In den meisten Lehrbüchern zur Testkonstruktion werden jedoch noch nicht einmal Methoden zur Überprüfung dieser sehr strengen Voraussetzungen besprochen. Daher kann es kaum verwundern, dass diese Annahmen in der Praxis äußerst selten überprüft werden. Wären sie tatsächlich erfüllt, so entfiele zudem jegliche Rechtfertigung für die übliche Selektion der Items anhand ihrer Trennschärfekoeffizienten, da die Trennschärfen der verschiedenen Items in diesem Fall identisch sein müssten.

Aber selbst bei nicht parallelen Items ist es äußerst fragwürdig, ob Trennschärfeanalysen dazu geeignet sind, die Reliabilität und Validität einer Skala sicherzustellen. Die meisten Lehrbücher zur Testtheorie befassen sich gar nicht mit dieser Frage. Sie vermitteln allenfalls durch die Diskussion des Zusammenhangs der Trennschärfe mit der Itemschwierigkeit implizit den Eindruck, dass die Ermittlung der Trennschärfe bei der Testkonstruktion von Nutzen ist. Allerdings legen die bekannten Zusammenhänge der Trennschärfe mit den Testgütekriterien eher

[21] Meist lässt sich eine deutliche Verkürzung der Skala ohne Einbußen bei der Validität erreichen.

nahe, auf trennscharfe Items bei der Testkonstruktion zu verzichten (vgl. [2.2-1] auf S. 30). Krauth (1995) empfiehlt jedoch, aufgrund der eingeschränkten Aussagekraft dieser Formeln auf Trennschärfeanalysen bei der Testkonstruktion ganz zu verzichten, da der Zusammenhang der Trennschärfe mit den Testgütekriterien unklar sei.

Tatsächlich zeigt sich bei einer eingehenden theoretischen Analyse der Beziehungen zwischen der Trennschärfe und den Testgütekriterien (Yousfi, 2004b, vgl. S. 26), dass die Präferenz für trennscharfe Items durchaus geeignet ist, die Reliabilität eines Tests sicherzustellen. Zudem wird (bei Skalen mit positiver Validität) die Auswahl von Items mit hoher Validität begünstigt. Dennoch kann man nicht erwarten, dass die Selektion anhand der Trennschärfe auch zu validen Skalen führt, da hierbei auch die Kumulation von kriteriumsirrelevanten Varianzanteilen (Fehlerredundanz, vgl. Kapitel 3.2) begünstigt wird.

Die theoretischen Analysen von Yousfi (2004a, 2004b) wurden durch die Ergebnisse der Simulationsstudien (Kapitel 5) und der empirischen Studien (Kapitel 6) weitgehend bestätigt. Dabei zeigte sich, dass die herkömmlichen, klassischen Methoden der Testkonstruktion allenfalls bei τ-kongenerischen Skalen angemessen sind. Die τ-Kongenerität stellt eine empirisch überprüfbare Hypothese über die Zusammenhänge der Itemantworten mit den Ausprägungen auf der latenten Variable dar. Derartige Hypothesen sind das zentrale Merkmal von Modellen der Item-Response Theorie. Auch in der klassischen Testtheorie werden solche Annahmen gemacht, wenngleich sie nicht immer explizit genannt werden. So setzen die Spearman-Brown Formel und alle gängigen Methoden zur Ermittlung der Reliabilität die essentielle τ-Äquivalenz sowie die Varianzhomogenität und Unkorreliertheit der Fehler voraus. All diese Annahmen lassen sich durchaus überprüfen (vgl. Steyer & Eid, 1993). Es scheint jedoch das zentrale Merkmal klassisch orientierter Testkonstruktion zu sein, dass solche Annahmen einfach vorausgesetzt werden, ohne dass sie, wie bei Modellen der Item-Response Theorie, einer empirischen Prüfung unterzogen werden. Daher stellt sich die Frage, ob die übliche Praxis der Testkonstruktion nicht treffender als naive Testkonstruktionsmethode zu bezeichnen wäre, die im Gegensatz zu einem testtheoretisch fundierten Vorgehen auf die Überprüfung von Modellannahmen verzichtet.

Die Modellannahmen klassischer und probabilistischer Testmodelle beziehen sich in der Regel auf die Eindimensionalität des Items sowie auf das Fehlen von statistischen Zusammenhängen

zwischen den Fehlerkomponenten der Items[22]. Diese Annahmen sind jedoch so restriktiv, dass sie in der Praxis häufig nicht erfüllt sind. Während die Mehrzahl der Testkonstrukteure daher lieber auf die Überprüfung derartiger Modellannahmen verzichtet, empfehlen Cattell und Tsujioka (1964) aus der Not eine Tugend zu machen und die Mehrdimensionalität des Itempools nicht nur zu tolerieren, sondern für die Testkonstruktion nutzbar zu machen. Der Skalensummenwert kann nämlich auch dann ein sinnvolles Messinstrument für eine bestimmte Eigenschaft sein, wenn die einzelnen Items auch andere Merkmale erfassen, sofern sich die Einflüsse der konfundierenden Merkmale auf die verschiedenen Items gegenseitig aufheben. Während Cattell und Tsujioka (1964) versuchen, die Ladungen auf fremden Faktoren auszubalancieren, um die Suppression dieser irrelevanten Varianzanteile zu erreichen, wurde in der vorliegenden Arbeit die Korrelation mit einem externen Kriterium maximiert.

Während das Vorgehen von Cattell und Tsujioka jedoch explizit an die Faktorenanalyse gebunden ist und darauf abzielt die Faktortreue des Skalensummenwerts sicherzustellen, versucht der in dieser Arbeit realisierte Ansatz systematische und unsystematische Fehler gleichermaßen zu minimieren. Die in der vorliegenden Arbeit verwendeten multivariaten Selektionsmethoden sind nicht an ein bestimmtes statistisches Verfahren oder Modell gebunden und können zur Optimierung beliebiger statistischer Kennwerte eingesetzt werden. Die in den Kapiteln 2.2 und 4.1 beschriebenen Algorithmen lassen sich beispielsweise auch dazu verwenden, um die Reliabilität und Validität bei einem Rasch-homogenen Itempool zu optimieren. Dazu nimmt man, statt der externen Validität, die Werte der Informationsfunktion (Rost, 1996) als Kriterium. Auch wenn man nicht die externe Validität der Skala, sondern die Faktortreue sensu Cattell und Tsujioka (1964) optimieren will, kann man die beschriebenen Algorithmen einsetzen. Dazu muss man lediglich die minderungskorrigierten Strukturkoeffizienten (=Item-Faktor Korrelationen) der Items als Itemvalidität betrachten und die Faktortreue anhand der Kovarianzmatrix der Items optimieren. Dabei werden die Kommunalitäten der Items als Reliabilitätsschätzungen verwendet.

Wenn die entsprechenden statistischen Kennwerte fehlerfrei ermittelt worden sind und das Evaluationskriterium mit dem Zielkriterium des Algorithmus übereinstimmt, kann man davon

[22] Unkorrelierte Messfehler werden in der Statistik häufig postuliert. Gerade bei der Testkonstruktion gibt es jedoch viele Bedingungen, die dazu führen, dass Messfehler korrelieren: (a) States beeinflussen die Bearbeitung aller Items eines Trait oder Leistungstests, (b) Tendenzen zur konsistenten Selbstdarstellung in Persönlichkeitsfragebögen, (c) reaktionskontingentes Lernen in Leistungstests etc..

ausgehen, dass die multivariaten Verfahren bessere Ergebnisse erzielen als die Selektion anhand von Itemkennwerten. Die Ergebnisse der Simulationsstudien und der empirischen Studien zeigen jedoch, dass die multivariaten Verfahren nicht unbedingt besser abschneiden als die Selektion anhand von Itemkennwerten, wenn Stichprobenfehler bei der Schätzung der statistischen Kennwerte nicht zu vernachlässigen sind. Welcher der in dieser Arbeit untersuchten multivariaten Selektionsalgorithmen zur Maximierung der Validität eingesetzt wird, scheint dabei keine Rolle zu spielen. Die Optimierung des MAXVAL-Verfahrens durch Vertauschungen von Items sowie die vollständige Permutation zeigten in keiner der Studien wesentlich andere Ergebnisse als das einfache MAXVAL-Verfahren von Ulrich (1985).

Erwartungsgemäß ist die Regression zur Mitte bei den multivariaten Verfahren besonders ausgeprägt, wenn nur ein kleiner Teil der Items aus einem umfangreichen Itempool ausgewählt wird, und wenn die Schätzfehler im Vergleich zu den wahren Unterschieden zwischen den zur Auswahl stehenden Itemkombinationen groß sind. Das Ausmaß des Schätzfehlers hängt vor allem vom Stichprobenumfang ab, während die wahren Unterschiede zwischen den Itemkombinationen von den Kommunalitäten der Items und des Kriteriums abhängen.

Multivariate Methoden zur Optimierung der Validität scheinen allenfalls bei einem hinreichend reliablen Kriterium, großen Personenstichproben, einem kleinen Itempool oder bei Elimination von wenigen Items aus einem umfangreichen Itempool zu valideren Skalen zu führen als die Selektion anhand der Itemvalidität. Die Selektion anhand der Itemvalidität scheint dabei jedoch grundsätzlich zu reliableren Skalen zu führen als die multivariate Optimierung der Validität. Dies spricht allerdings nicht unbedingt für die Verwendung der Itemvalidität als Selektionskriterium, da eine höhere Reliabilität bei gleicher oder geringerer Validität impliziert, dass der systematische Fehler (Bias) bei der Erfassung des Merkmals größer ist (vgl. [1.2-6] auf S. 18). Dies liegt daran, dass die Selektion anhand der Itemvalidität lediglich zur Kumulation valider Varianz führt, während bei den multivariaten Verfahren auch Suppressionseffekte zur Optimierung der Validität genutzt werden. Bei den Selektionsmethoden, die eine Homogenisierung der Skala anstreben, kumuliert dagegen sowohl die valide als auch die systematische Fehlervarianz. Da die systematische Fehlervarianz wahre Varianz im Sinne der klassischen Testtheorie ist, wird daher trotz hoher Reliabilität oft nur eine mäßige Validität erreicht.

Systematische Fehler führen bei der praktischen Anwendung einer Skala zu größeren Problemen, da sie im Gegensatz zu unsystematischen Fehlern nicht nur die Präzision der Aussagen beeinträchtigen, sondern auch zu Fehlinterpretationen und falschen Schlussfolgerungen führen[23]. Dies ist letztlich auch der Grund dafür, dass die Validität als übergeordnetes Gütekriterium und finales Ziel der Testkonstruktion betrachtet werden kann, während eine hinreichende Reliabilität eher ein instrumentelles Ziel und eine notwendige Bedingung für eine valide Skala ist.

Andererseits stellt die Heterogenität der Skalen, die bei multivariater Optimierung der Reliabilität resultieren, ein Problem für die Schätzung der Reliabilität dar. In den vorgestellten Simulationsstudien musste die Reliabilität gar nicht geschätzt werden, da sie sich anhand der Populationsparameter der Items berechnen ließ. In der Empirie beruhen die meisten Methoden zur Schätzung der Reliabilität jedoch auf der Homogenität einer Skala. Bei heterogenen Skalen wird die Reliabilität daher deutlich unterschätzt. Dies hat wiederum zur Folge, dass die (minderungskorrigierte) konvergente bzw. divergente Validität mit anderen Tests überschätzt wird. Daraus ergibt sich die Gefahr, dass Unterschiede im Geltungsbereich verschiedener Tests nicht erkannt werden. Bei multivariater Konstruktionsmethodik sollte die Reliabilität daher bevorzugt durch Retest-Korrelationen geschätzt werden.

Es hängt jedoch letztlich vom Forschungsansatz ab, welche Bedeutung man der Validität, Reliabilität und Homogenität einer Skala beimisst. Bei einem induktiven Vorgehen geht es darum, Wissen über die statistischen Beziehungen zwischen verschiedenen Persönlichkeitsmerkmalen zu sammeln und daraus Theorien abzuleiten. In diesem Kontext dürfte es häufig schwer fallen, externe Variablen zu finden, die echte Kriterien im Sinne von Burisch (1984a) sind und nicht nur Quasi-Kriterien oder Zielvariablen (vgl. Kapitel 1.2.2, S. 18). Die Maximierung der externen Validität dürfte daher nicht das zentrale Ziel der Testkonstruktion sein. Die Bildung von homogenen Skalen erleichtert in dieser Phase des Forschungsprozesses die inhaltliche Interpretation der Testwerte und sichert gleichzeitig eine hohe Reliabilität sowie präzise Reliabilitätsschätzungen. Dies ermöglicht die Klärung von Beziehungen zu anderen Konstrukten. Bei induktiver Forschungsstrategie bieten sich zudem operationale Definitionen des Messgegenstands an. In diesem Fall ist die Validität eines Tests äquivalent zur Reliabilität.

[23] Bei systematischen Fehlern besteht im Prinzip die Möglichkeit sie durch Korrekturformeln auszugleichen oder sie einfach bei der Interpretation der Testergebnisse zu berücksichtigen. Dies setzt jedoch voraus, dass man die Natur des systematischen Fehlers kennt. In diesem Fall sollte es jedoch möglich sein, den Test so zu konstruieren, dass die systematischen Fehler erst gar nicht auftreten (z.B. durch die Ausnutzung von Suppressionseffekten).

Eine rein induktive Forschungsstrategie wird jedoch „unter den Wissenschaftstheoretikern ... nur mehr von den provinziellsten und ungebildedsten Geistern ernst genommen“ (Lakatos, 1982, S. 173). Induktion gilt zwar durchaus als *ein* probates Mittel, um Hypothesen über den Forschungsgegenstand zu generieren und Informationen zu sammeln. In der Regel wird es jedoch im Zuge der Theorienbildung zu einer Redefinition des Messgegenstands kommen. In der Folge treten dann meist Mängel des ursprünglichen Tests auf. Letztlich führt daher kein Weg an der Überprüfung der gewonnenen Hypothesen vorbei. Die aufgestellten Hypothesen über die statistischen Zusammenhänge eines Tests mit externen Variablen sollten sich aus den zugrundegelegten Persönlichkeitstheorien ableiten lassen. Nur wenn nach Maßgabe dieser Theorien eine externe Variable vorhanden ist, die den Status eines echten Kriteriums im Sinne von Burisch (1984a) hat, ist die Optimierung der externen Validität als Testkonstruktionsmethode angezeigt. In diesem Fall sollte die (externe) Validität zentrales Anliegen der Testkonstruktion sein. Eine hohe Validität bedingt zudem eine hohe Reliabilität. Die Homogenität des Tests ist daher von untergeordneter Bedeutung.

Falls jedoch keine externe Variable vorhanden ist, die den Status eines echten Außenkriteriums hat, dann sollte eine Messvorschrift entwickelt werden, die nach Maßgabe der zugrundegelegten Theorie geeignet sein sollte, das entsprechende Persönlichkeitsmerkmal zu erfassen. In der Regel besteht in der Psychologie die Messvorschrift in der Vorgabe der Items eines Persönlichkeitstests. Aus der inhaltlichen Theorie sollte nach Möglichkeit abzuleiten sein, welches statistische Testmodell angemessen ist. Ist dies nicht der Fall, so weist dies auf einen Mangel an Präzision der inhaltlichen Theorie hin. In diesem Fall sollte ein Testmodell gewählt werden, das möglichst wenig Voraussetzungen hat, z.B. das ordinale Mokken-Modell. Wenn selbst liberale Testmodelle nicht im Einklang mit den erhobenen Daten stehen, dann stellt dies eine empirische Widerlegung der zugrundegelegten Theorie dar. Daher reicht es nicht, Items zu eliminieren, die nicht im Einklang mit den Modellannahmen stehen, ohne zu klären, warum es zu den Verletzungen der Modellannahmen kam. Falls sich Störungshypothesen nicht empirisch bestätigen lassen, sollte eine Revision der Theorie oder zumindest eine Einschränkung des Geltungsbereichs in Betracht gezogen werden.

Das eben skizzierte Vorgehen stellt sicherlich den Idealfall der theoriegeleiteten, deduktiven Testkonstruktion dar. Der von Burisch (1984a) geprägte Begriff der deduktiven Skalenkonstruktion ist dagegen wesentlich breiter. Er umfasst alle Methoden, bei denen implizite oder explizite theoretische Vorstellungen Grundlage für die Formulierung der Testitems sind. Insbesondere im Persönlichkeitsbereich wird jedoch der Analyse der psychologischen Prozesse,

die dem Verhalten der Testperson in der Testsituation zugrunde liegen, zuwenig Aufmerksamkeit geschenkt. Stattdessen wird einfach aus der Markierung von vorgegebenen Antwortalternativen auf Verhaltenstendenzen geschlossen, die sich auf Situationen beziehen, die mit der Testsituation oft nichts gemeinsam haben.

Aber selbst wenn man die Fähigkeit und die Bereitschaft von Testpersonen zu validen Selbstberichten weniger skeptisch beurteilt, sollte man sich bei der Itemselektion keinesfalls zu sehr von den statistischen Kennwerten der entsprechenden Items leiten lassen. Items, die man aus theoretischen Gründen für gut hält, sollte man also keinesfalls allein wegen einer geringen Trennschärfe eliminieren. Eine geringe Trennschärfe stellt nämlich in der Regel weder eine Verletzung von Modellannahmen dar, noch ist sie ein Indikator dafür, dass sich das Item negativ auf die Validität des Tests auswirkt. Es muss daher ernsthaft in Zweifel gezogen werden, ob die üblichen Trennschärfeanalysen sich tatsächlich positiv auf die Qualität eines Tests auswirken. Die resultierenden Tests sind häufig nicht valider als bei zufälliger Auswahl. Sie begünstigen die Auswahl von Items mit systematischen und korrelierten Fehlerkomponenten.

Wenn man tatsächlich die einzelnen Items als Messinstrumente für die betreffende Persönlichkeitseigenschaft ansieht, dann führt an Modelltests, wie sie in der Item-Response Theorie üblich sind, kein Weg vorbei. Verzichtet man jedoch auf die Annahme der Eindimensionalität der Items, dann sollte man multivariate Methoden zur Optimierung der Validität als Testkonstruktionsmethode in Betracht ziehen, wenn ein sinnvolles externes Kriterium verfügbar ist. Sie führen häufig zu validen Tests und minimieren dabei vor allem systematische Fehler durch die Ausnutzung von Suppressionseffekten. Wenn man aufgrund einer kleinen Personenstichprobe, unreliablen Items oder eines unreliablen Kriteriums Zweifel an der Angemessenheit der multivariaten Verfahren hat oder wenn nur ein kleiner Teil eines umfangreichen Itempools in den Test aufgenommen werden soll, sollte man die Itemvalidität als Selektionskriterium verwenden. In jedem Fall sollte man nie vergessen, dass die Messung psychischer Eigenschaften fundiertes Wissen über den Gegenstand der Messung voraussetzt und keine Dienstleistung ist, die man an Statistiker delegieren könnte.

8 Zusammenfassung

In der vorliegenden Arbeit wird der Nutzen von multivariaten Methoden der Itemselektion untersucht. Im Gegensatz zu herkömmlichen Methoden zur Konstruktion psychologischer Tests anhand von Itemkennwerten berücksichtigen diese Methoden, dass die Wirkung eines Items auf die Gütekriterien des Tests nicht unabhängig davon beurteilt werden kann, welche Items außerdem im Test enthalten sind. Ausgehend von einer Darstellung der klassischen Testtheorie, die streng zwischen tautologischen Aussagen und Sätzen mit empirischem Gehalt differenziert, wird gezeigt, dass multivariate Methoden der Skalenkonstruktion auf der Nutzbarmachung von Suppressionseffekten beruhen. Es wird eine formale Definition von Suppression in der Testkonstruktion präsentiert und deren Nutzen bei der Interpretation von empirischen Daten demonstriert.

Außerdem werden im empirischen Teil der Arbeit die diskutierten Itemselektionsmethoden sowohl durch Monte-Carlo Simulationen als auch anhand empirischer Daten verglichen. Es zeigt sich, dass multivariate Methoden der Skalenkonstruktion nur dann zu valideren Tests führen als herkömmliche Methoden, wenn der Stichprobenumfang groß, der Itempool mehrdimensional oder die Kommunalität der Kriteriums groß ist. Ansonsten führen multivariate Methoden nur in der Stichprobe zu einer besonders hohen Validität, während sie in der Population vor allem bei der Auswahl eines geringen Teils der Items zu wenig validen Skalen führen. Außer bei einem eindimensionalen Itempool führen multivariate Methoden der Itemselektion dagegen zu Tests von vergleichsweise geringer Reliabilität. Die Selektion anhand der Trennschärfe führt bei mehrdimensionalem Itempool dagegen nur bei der Auswahl von wenigen Items zu Tests mit befriedigender Validität, während bei der Aufnahme eines Großteils der Items häufig sogar weniger valide Skalen resultieren als bei zufälliger Auswahl. Bei Optimierung von Cronbach's α resultieren, außer bei eindimensionalem Itempools, Skalen mit hoher Reliabilität, aber mit sehr geringer Validität. Die Selektion anhand der Itemvalidität führt dagegen unabhängig von den Eigenschaften des Itempools zu reliablen und validen Skalen. Die Sicherung der Itemvalidität kann daher als einzige der untersuchten Methoden unabhängig von den Eigenschaften des Itempools empfohlen werden, während die anderen Verfahren nur dann angewendet werden sollten, wenn die entsprechenden Voraussetzung erfüllt sind.

9 Literatur

Amelang, M. & Bartussek, D. (2001). *Differentielle Psychologie und Persönlichkeitsforschung.* Stuttgart: Kohlhammer.

Amelang, M. & Borkenau, P. (1981). Über die faktorielle Struktur und externe Validität einiger Fragebogen-Skalen zur Erfassung von Dimensionen der Extraversion und emotionalen Labilität. *Zeitschrift für Differentielle und Diagnostische Psychologie, 3* (2), 119-146.

Amelang, M. & Zielinski, W. (1997). *Psychologische Diagnostik und Intervention (2. Aufl.).* Berlin: Springer.

Amelang, M. & Zielinski, W. (2001). *Psychologische Diagnostik und Intervention (3. Aufl.).* Berlin: Springer.

Amelang, M., Schäfer, A. & Yousfi, S. (2002). Comparing verbal and non-verbal personality scales: Investigating the reliability and validity, the influence of social desirability, and the effects of fake good instructions. *Psychologische Beiträge, 44(1)*, 24-41.

Anderson, R. D., & Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability, 5*, 111-150.

Bartlett, M. S. (1937). the statistical conception of mental factors. *British Journal of Psychology, 28,* 97-104.

Basilevsky, A. (1994). *Statistical Factor Analysis and related methods.* New York: Wiley.

Bell, R. & Lumsden, J. (1980). Test length and validity. *Applied Psychological Measurement, 4, 165-170.*

Berk, K.N.(1978). Comparing subset regression procedures. *Technometrics, 20(1),* 1-6.

Birnbaum, A. (1968). Some latent trait models an their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores (pp. 397-479)*. Reading, Massachusetts: Addison-Wesley.

Borkenau, P. (1981). *Individuelle Variabilität und differentielle Vorhersagbarkeit.* Unveröffentlichte Dissertation: Ruprecht-Karls-Universität Heidelberg.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler (5. Aufl.).* Berlin: Springer.

Broughton, R. (1984). A prototype strategy for construction of personality scales. *Journal of Personality and Social Psychology, 47,* 1334-1347.

Bryson, R. (1972). Shortening tests: Effects of method used, length, and internal consistency on correlation with total score. *Proceedings of the Annual Convention of the American Psychological Association, Vol. 7(Pt. 1)*. 7-8.

Burisch, M. (1984a). Approaches to personality inventory construction. A comparison of merits. *American Psychologist, 39(3),* 214-227.

Burisch, M. (1984b). You don't always get what you pay for: Measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality, 18,* 81-98.

Burisch, M. (1997). Test-length and validity revisited. *European Journal of Personality, 11,* 303-315.

Campbell, G.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cattell, R.B. & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement, 24(1),* 3-30.

Cattell, R.B.& Radcliffe, J.A. (1962). Reliabilities and validities of simple and extended weighted and buffered unifactor scales. *The British Journal of Statistical Psychology, 15(2)* 113-127.

Cohen, J. & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, J. (1988). *Statistical Power analysis for the behavioral sciences (2nd ed.).* New York: Erlbaum.

Conger, A.J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement, 34,* 35-46.

Darlington, R. B. & Bishop, C.H. (1966). Increasing test validity by considering interitem correlations. *Journal of Applied Psychology, 50(4),* 322-330.

Darlington, R.B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 69,* 161-182.

Diehr, G. & Hoflin, D.R. (1974). Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics, 16(2),* 317-320.

Edwards, R. H. (1981) Coefficients of effective length. *Educational and Psychological Measurement, 41,* 283-285.

Fahrenberg, F., Selg, H. & Hampel, R. (1973). *Das Freiburger Persönlichkeitsinventar.* Göttingen: Hogrefe.

Fischer, G. (1974). *Einführung in die Theorie psychologischer Tests.* Bern: Verlag Hans Huber.

Fischer, G. (1993). *Lineare Algebra (9. Aufl.)*. Braunschweig: Vieweg.

Fisz, M. (1966). *Wahrscheinlichkeitsrechnung und mathematische Statistik.* Berlin: VEB Deutscher Verlag der Wissenschaften.

Forster, O. (1999). *Analysis 1* (5. Aufl.). Braunschweig: Vieweg.

Furnival, G.M. & Wilson, R.W.,Jr. (1974). Regression by leaps and bounds. *Technometrics, 16,* 499-512.

Gigerenzer, G. (1981). *Messung und Modellbildung in der Psychologie.* München: E. Reinhardt.

Gleser, G.C. & Dubois, P.H. (1951). A successive approximation method of maximizing test validity. *Psychometrika, 16(1),* 129-139.

Graybill, F.A. (1961). *An introduction to linear statistical models.* New York: McGraw-Hill.

Green, B.F., Jr. (1954). A note on item selection for maximum validity. *Educational and Psychological Measurement, 14,* 161-164.

Greif, S. (1970). Untersuchung zur deutschen Übersetzung des 16 PF-Fragebogens. *Psychologische Beiträge, 2,* 186-213.

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

Guttman, L. (1945). A basis for analysing test-retest reliability. *Psychometrika, 10,* 255-282.

Harwell, M.R. (1997). Analyzing the results of Monte Carlo studies in item response theory. *Educational and Psychological Measurement, 57(2),* 266-279.

Heisenberg, W. & Bohr, N. (1963). *Die Kopenhagener Deutung der Quantentheorie*. Stuttgart: Battenberg.

Holling, H. (1981a). Das Suppressorkonzept: Eine systematische Analyse und Neudefinition. *Zeitschrift für Differentielle und Diagnostische Psychologie, 2(2),* 123-150.

Holling, H. (1981b). Homogenität und faktorielle Validität von Skalen. *Diagnostica, 27(2),* 97-106.

Holling, H. (1983). Suppressor structures in the general linear model. *Educational and Psychological Measurement, 43,* 1-9.

Horst, P. (1936). Item selections by means of a maximizing function. *Psychometrika, 1,* 229-244.

Horst, P. (1941). The role of prediction variables which are independent of the criterion. In P. Horst (Ed.): The prediction of personal adjustment. *Social Science Research Bulletin, 48,* 431-436.

Howarth, E., Browne, J.A. & Marceau, R. (1972). An item analysis of Cattell's 16 PF. *Canadian Journal of Behavioural Science, 4,* 85-90.

Jackson, D.N. (1984). *Personality Research Form manual.* Port Huron, MI: Research Psychologists Press.

Jäger, A.O.; Süss, H.M. & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test.* Göttingen: Hogrefe.

Kendall, M.G. & Stuart, A. (1977). *The advanced theory of statistics (Vol.1): Distribution Theory (4th ed.).* New York: Hafner.

Krauth, J. (1995). *Testkonstruktion und Testtheorie*. Weinheim: PVU.

Lakatos, I. (1982). *Die Methodologie der wissenschaftlichen Forschungsprogramme (Philosophische Schriften Bd. 1)*. Braunschweig: Vieweg.

Laplace, P. S. (1814). *Essai philosophique sur le probabilités*. Paris: Gauthier-Villars.

Lienert, G.A.& Raatz, U. (1994). *Testaufbau und Testanalyse.* Weinheim: PVU.

Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996), *SAS System for Mixed Models,* Cary, NC: SAS Institute Inc.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51,* 493-504.

Lord, F.M. (1955). Some perspectives on "the attenuation paradox" in test theory. *Psychological Bulletin, 52,* 505-510.

Lord, F.M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48(2),* 233-245.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley.

Lubin, A. (1957). Some formulae for use with suppressor variables. *Educational and Psychological Measurement, 17,* 286-296.

McFatter, R.M. (1979). The use of structural equations models in interpreting regression equations including suppressor and enhancer variables. *Applied psychological measurement, 3(1),* 123-135.

Meyer, A.E., Arnold, M.-A., Freitag, D.E. & Balck, F. (1977). Cattells Test-Konstruktions-Strategie, beurteilt an der Eppendorf-Übersetzung seines 16 Persönlichkeits-Faktoren (16 PF)-Fragebogens. *Diagnostica, 25,* 1041-1048.

Moosbrugger, H. & Zistler, R. (1993). Wie befreit man die Trennschärfe von den Zwängen der Item-Schwierigkeit? Das SPS-Verfahren. *Diagnostica, 19,* 22-43.

Moosbrugger, H. (1988). Testtheorie: Klassische Ansätze. In R.S. Jäger (Hrsg.), *Psychologische Diagnostik: Ein Lehrbuch.* (S. 253-264). München: PVU.

Müller, H. & Moosbrugger, H. (1985). Zur Bestimmung von Konfidenzintervallen für nicht-normierte und normierte „wahre Werte". *Diagnostica, 31, 279-288.*

Müller, H. (2000). Summenscore und Trennschärfe beim Rasch-Modell. *Psychologische Rundschau. 51(1),* 34-35.

Muller, K. E., LaVange, L. M., Ramey, S. L. & Ramey, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association, 87,* 1209-1226.

O'Brien, R.G. & Kaiser, M.K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin, 97,* 316-333.

Odell, P.L. & Feiveson, A.H. (1966) A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association,61,* 199-203.

Olson, C.L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin, 83*, 579-586.

Paunonen, S.V. & Jackson, D.N. (1985). The validity of formal and informal Personality Assessments. *Journal of Research in Personality, 19,* 331-342.

Paunonen, S.V. (1984). Optimizing the validity of personality assessments: The importance of aggregation and item content. *Journal of Research in Personality, 18,* 411-431.

Reinsch, C.H. (1967). Smoothing by Spline Functions. *Numerische Mathematik, 10,* 177-183.

Rost, J. & von Davier, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement, 18(2),* 171-182.

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

Rost, J. (1996). Lehrbuch Testtheorie Testkonstruktion. Bern: Huber.

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau. 50(3),* 140-156.

Rost, J. (2000). Haben ordinale Rasch-Modelle variierende Trennschärfen? *Psychologische Rundschau. 51(1),* 36-37.Rost, J. (2000).

Samejima, F. (1993). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika 58,* 119-138.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18(3),* 229-244.

SAS/IML Software: Usage and Reference Version 6 (1st ed.) (1989). Cary, NC: SAS Institute.

Saville, P. & Blinkhorn, S. (1981). Reliability, homogeneity and the construct validity of Cattell's 16 PF. *Personality and Individual Differences, 2,* 325-333.

Searle, S.R., Speed, R.M. & Milliken, G.A. (1980). Populations Marginal Means in the Linear Model: An alternative to testing Least Squares Means. *The American Statistician, 34*, 216-221.

Smith, R.L.; Ager, J.W. & Williams, D.L.; (1992). Suppressor variables in multiple regression/correlation. *Educational and Psychological Measurement, 52(1),* 17-29.

Stern, I. & Tzelgov, J. (1978). Comments on two statements about three-variate multiple regression. *Psychological Reports, 43,* 687-690.

Stevens, J.P. (1979). Comment on "Choosing a test statistic in multivariate analysis" by C.L. Olson. *Psychological Bulletin, 88,* 728-737.

Stevens, J.P. (1980). Power of the multivariate analysis of variance tests. *Psychological Bulletin, 88,* 728-737.

Steyer, R. & Eid, M. (1993). *Messen und Testen.* Berlin: Springer.

Thompson, M.L. (1978). Selection of variables in multiple regression: Part I. A review and evaluation. *International Statistical Review, 46,* 1-19.

Thompson, M.L. (1978). Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples . *International Statistical Review, 46,* 129-146.

Thurstone, L. L. (1935). *The vectors of mind.* Chicago: University of Chicago Press.

Toops, H.A. (1941). The L-method. *Psychometrika, 6,* 249-266.

Tzelgov, J. & Henik, A. (1981). On the differences between Conger's and Velicer's definitions of suppressor. *Educational and Psychological Measurement, 41,* 1027-1031.

Tzelgov, J. & Henik, A. (1985). A definition of suppression situations for the general linear model: A regression weights approach. *Educational and Psychological Measurement, 45,* 281-284.

Tzelgov, J. & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin, 109(3),* 524-536.

Tzelgov, J. & Stern, I. (1978). Relationships between variables in three variables linear regression and the concept of suppressor. *Educational and Psychological Measurement, 38,* 325-335.

Ulrich, R. (1985). Die Beziehung zwischen Testlänge und Validität für nicht-parallele Aufgaben: Verschiedene Methoden der Validitätsmaximierung. *Zeitschrift für Differentielle und Diagnostische Psychologie, 6* (1), 32-45.

Velicer, W.F. (1978). Suppressor variables and the semipartial correlation coefficient. *Educational and Psychological Measurement, 38,* 953-958.

Webster, H. (1953). Maximizing test validity by item selection. *Psychometrika, 21,* 153-164.

Werner, J. (1997). *Lineare Statistik.* Weinheim: PVU.

West, S.G.; Aiken, L.S.; Krull, J.L..(1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality, 64(1)*, 1-48.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik. Ein Lehrbuch zur Psychologischen Methodenlehre.* Göttingen: Hogrefe.

Williams, R.H. & Zimmerman, D.W. (1982). Reconsideration of the "attenuation paradox"-and some new paradoxes in test validity. *Journal of Experimental Education, 50(3),* 164-171.

Willkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin, 86(1),* 168-174.

Yousfi, S. (1999). *Prädiktoren der Itemvalidität.* Unveröffentlichte Diplomarbeit: Universität Heidelberg.

Yousfi, S. (2004a). *Mythen und Paradoxien der klassischen Testtheorie (I): Testlänge und Gütekriterien.* Manuskript zur Veröffentlichung in der Diagnostica angenommen.

Yousfi, S. (2004b). *Mythen und Paradoxien der klassischen Testtheorie (II): Trennschärfe und Gütekriterien.* Manuskript zur Veröffentlichung in der Diagnostica angenommen.

Zimmerman, D.W. (1975). Two concepts of true score in test theory. *Psychological Reports, 36,* 795-805.

Zimmerman, D.W. & Williams, R.H. (1977). The theory of test validity and correlated orrors of measurement. *Journal of Mathematical Psychology, 16*, 135-152.

Zimmerman, D.W. & Williams, R.H. (1980). Is classical test Theory ‚robust' under violation of the assumption of uncorrelated Errors? *Canadian Journal of Psychology, 34*(3), 227-236.