Siegener Periodicum zur Internationalen Empirischen Literaturwissenschaft

Herausgegeben von Reinhold Viehoff (Halle/Saale)
Gebhard Rusch (Siegen)
Rien T. Segers (Groningen)

Jg. 18 (1999), Heft 1

Peter Lang Europäischer Verlag der Wissenschaften

SPIEL
Siegener Periodicum zur Internationalen Empirischen Literaturwissenschaft

SPIEL:
Siegener Periodicum zur Internationalen Empirischen Literaturwissenschaft

Jg. 18 (1999), Heft 1

Peter Lang
Frankfurt am Main · Berlin · Bern · Bruxelles · New York · Wien

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Siegener Periodicum zur internationalen empirischen Literaturwissenschaft (SPIEL)
Frankfurt am Main ; Berlin ; Bern ; New York ; Paris ; Wien : Lang
ISSN 2199-8078
Erscheint jährl. zweimal

JG. 1, H. 1 (1982) -
[Erscheint: Oktober 1982]

NE: SPIEL

ISSN 2199-8078

Siegener
Periodicum zur
Internationalen
Empirischen
Literaturwissenschaft

SONDERHEFT / SPECIAL ISSUE SPIEL 18 (1999), H.1

Special Issue/Sonderheft

The methodology of interdisciplinary research on reading and socialization.

Interdisziplinäre Methodik der Lesesozialisationsforschung.

ed. by/hrsg. von
Norbert Groeben (Köln)

Siegener
Periodicum zur
Internationalen
Empirischen
Literaturwissenschaft

Inhalt / Contents SPIEL 18 (1999), H.1

10.3726/80984_21

SPIEL 18 (1999), H. 1, 21–34

Johannes Naumann, Tobias Richter & Ursula Christmann

Psychologische Tests und Fragebogen

In the social sciences, psychological tests and questionnaires are standard procedures of data collection which belong to the ‚quantitative paradigm'. First, we give a contrastive explication of the concepts ‚test' and ‚questionnaire'. Next, we sketch a content-related taxonomy of psychological tests and present an overview of psychological test theories. Measurement-theoretical foundations, psychological processes underlying respondents' behavior, social aspects of the testing situation, and selected ethical considerations are discussed as problems which concern tests and questionnaires to different degrees.

Dieser Beitrag soll eine kurze Einführung zu psychologischen Tests und Fragebogen als wichtigen Erhebungsmethoden in der Psychologie und den empirischen Sozialwissenschaften generell geben. Die illustrierenden Beispiele stammen vor allem aus dem Forschungsprojekt „Wechselwirkungen zwischen Verarbeitungsstrategien von traditionellen (linearen) Buchtexten und zukünftigen (nichtlinearen) Hypertexten" (vgl. Christmann, Groeben, Flender, Naumann & Richter 1999); es handelt sich dabei um Instrumente zur Erhebung potenzieller Kovariaten in einer experimentellen Untersuchung, und zwar zum einen um das Inventar zur Computerbildung (INCOBI, Richter, Naumann, & Groeben, im Druck), das vier Skalen zu Aspekten von Computer Literacy (zwei Wissenstests und zwei Selbsteinschätzungsskalen) und acht Skalen zur inhaltlich differenzierten Erfassung computerbezogener Einstellungen enthält. Zum anderen wurde ein computergestütztes Instrument zur Erfassung von Lesefähigkeiten entwickelt, dessen Skalen die Bewältigung von Verstehensstrategien (sensu van Dijk & Kintsch 1983) erfassen sollen, sowie ein Instrument, das Selbsteinschätzungen zur Bewältigung verschiedener Schreibstrategien erfragt. Schließlich erheben wir über einen mehrdimensionalen Test das Wissen der Probanden/innen zum Inhaltsbereich ‚Visuelle Wahrnehmung'.

1. Begriffsklärungen und Arten psychologischer Tests

Psychologischen Tests und Fragebogen ist ihre Zugehörigkeit zum ‚quantitativen Paradigma' gemeinsam (s.o. Beitrag Dzeyk & Groeben), was sich vor allem in dem methodischen Grundsatz einer hohen Standardisierung von Datenerhebungs- und Auswertungssituation widerspiegelt. Teilweise wird der Begriff ‚Test' nur auf objektive, d.h. ihrem Anspruch nach verfälschungssichere Verfahren zur Erfassung kognitiver Fähigkeiten angewandt, während den Verfahren, die auf subjektiven Selbstauskünften der Befragten beruhen, der Begriff ‚Fragebogen' vorbehalten bleibt (z.B. Eysenck 1953; Mummendey

1987). In diesem Beitrag gehen wir dagegen von einer alternativen Begriffsexplikation aus, die ‚Test' und ‚Fragebogen' anhand ihrer methodischen Fundierung, diagnostischen Zielsetzung und den darin enthaltenen theoretischen Implikationen unterscheidet (vgl. ähnlich Bortz & Döring 1995, 231).

Unter psychologischen *Tests* verstehen wir standardisierte und routinemäßig anwendbare Verfahren, mit denen die (in der Regel kontinuierliche) Ausprägung einer bestimmten, direkter Beobachtung nicht zugänglichen Personeneigenschaft erfasst werden soll (vgl. Lienert & Raatz 1994). Das Grundprinzip psychologischen Testens besteht dabei in der gezielten Vorgabe einer Menge standardisierter Reize (Testitems), die so zusammengestellt sind, dass die Reaktionen der getesteten Personen als Indikatoren für die Ausprägung des nicht direkt beobachtbaren (latenten) Merkmals gelten können. Als ein Beispiel kann der Wissenstest zu deklarativem Computerwissen (TECOWI) aus dem Inventar zur Computerbildung (INCOBI) dienen. Die Aufgaben des Wissenstests TECOWI, mit denen die Bedeutung von zwölf Begriffen aus dem Computerbereich erfragt wird, sind so ausgewählt, dass ihre Lösungswahrscheinlichkeit von der Ausprägung des Merkmals ‚Umfang des deklarativen Computerwissens' abhängen soll. Je stärker dieses Merkmal bei einer Person ausgeprägt ist, desto mehr Aufgaben des TECOWI sollte sie lösen können. Entsprechend wird von der Anzahl richtig gelöster Aufgaben auf die Ausprägung des Merkmals ‚Umfang des deklarativen Computerwissens' geschlossen. Dieser Schluss ist nur dann gerechtfertigt, wenn die Konstruktion des Tests nach bestimmten Regeln erfolgt ist und der Test bestimmte Kriterien erfüllt, die im Rahmen der *Testtheorie* spezifiziert werden (s. Abschnitt 3).

Fragebogen sind dagegen in der Regel nicht zur Erfassung latenter psychologischer Merkmale konzipiert, sondern eher als eine standardisierte und ökonomische Form der schriftlichen Befragung anzusehen. Das Ziel von Fragebogen in diesem Sinne ist die Erhebung objektiver Daten, z.B. der Häufigkeit konkreter Verhaltensweisen, oder die Erfassung konkreter Verhaltensintentionen (vgl. etwa die bekannte ‚Sonntagsfrage' der Meinungsforschungsinstitute). Alle vier im Rahmen unseres Projekts entwickelten Instrumente enthalten auch einen Fragebogen, mit dem Informationen zur tatsächlichen Computernutzung, zu Lese- und Schreibaktivitäten oder zur bisherigen Beschäftigung mit dem Thema ‚Visuelle Wahrnehmung' erfasst werden.

Psychologische Tests lassen sich nach einer Reihe von Aspekten klassifizieren, z.B. nach normierten und nicht-normierten, ein- und mehrdimensionalen, norm- oder kriteriumsorientierten Tests, nach der Art der Testdarbietung oder der vorgesehenen Antwortformate (vgl. Lienert & Raatz 1994). Ein normierter Test liegt vor, wenn Informationen über die Verteilung der Testwerte an einer hinreichend großen Referenzgruppe von Personen erhoben wurden, zu dem sich die Testwerte einer Person in Beziehung setzen lassen. Ein- und mehrdimensionale Tests werden nach der Anzahl der Merkmalsdimensionen oder Teilmerkmale unterschieden, die der Test über unterschiedliche Skalen erfassen soll. Bei kriteriumsorientierten (im Gegensatz zu normorientierten) Tests werden nicht Referenzgruppen von Personen, sondern Eigenschaften des Inhaltsbereich, auf den sich der Test bezieht, zur Interpretation der Testwerte herangezogen (z.B. Lernziel-Tests). An verschiedenen Darbietungsformen existieren Papier-Bleistift-Tests, materialbasierte Tests und seit einigen Jahren vermehrt computergestützte Tests. Bei den Antwortformaten der einzelnen Items werden auf Grund ihrer höheren Auswertungsobjektivität (s.o. Beitrag

Dzeyk & Groeben) Aufgaben mit gebundenem Antwortmodus solchen mit offenem Antwortmodus vorgezogen. Aufgaben mit gebundenem Antwortmodus können aus dichotomen Items (z.B. Richtig-Falsch-Aufgaben), Mehrfachwahl-Aufgaben (unter mehreren Antwortmöglichkeiten ist eine Variante auszuwählen), Zu- und Umordnungsaufgaben bestehen. Bei Selbst-Einschätzungsskalen (Rating-Skalen) werden entweder Stufenantwort-Aufgaben (z.B. mit den Antwortkategorien „stimme zu" – „neutral" – „stimme nicht zu") oder Aufgaben mit kontinuierlichem Antwortmodus verwendet (vgl. z.B. Lienert & Raatz 1994).

Unter inhaltlichen Gesichtspunkten, d.h. nach der Art der zu erfassenden psychologischen Merkmale, können zunächst zwei große Klassen psychologischer Tests unterschieden werden, *Leistungstests* und *Persönlichkeitstests* (Brickenkamp 1996; Amelang & Zielinski 1997). Leistungstests dienen der Erfassung kognitiver Leistungen, wobei zwischen allgemeinen Leistungstests (Intelligenz-, Vigilanz- und Konzentrationstests), speziellen Fähigkeitstests und Entwicklungstests unterschieden werden kann. Leistungstests mit einer besonderen diagnostischen Zielsetzung sind Berufs- oder Studieneignungstests sowie Schuleintrittstests. Persönlichkeitstests stellen Tests zur Erfassung von Persönlichkeitsmerkmalen im weiteren Sinn dar (ohne Aspekte kognitiver Leistungsfähigkeit). Demnach lassen sich unter Persönlichkeitstests sowohl Tests zur Erfassung persönlichkeitstheoretischer Konstrukte (z.B. Extraversion oder Neurotizismus) als auch Verfahren zur Erfassung von Interessen, von emotionalen und motivationalen Dispositionen und von Einstellungen subsumieren.

Die Skalen der im Rahmen unseres Projekts entwickelten Instrumente sind unterschiedlichen Testarten zuordnen. Das Inventar zur Computerbildung (INCOBI) enthält Wissenstests zur Erfassung deklarativen und praktischen Computerwissens (TECOWI und PRACOWI), die sich unter die speziellen Leistungstests subsumieren lassen. In dieselbe Klasse gehört das computergestützte und mehrdimensionale Instrument zur Erfassung von Lesefähigkeiten sowie der gleichfalls mehrdimensional konzipierte Vorwissenstest zu ‚Visueller Wahrnehmung'. Die Skalen des INCOBI, die Einstellungen gegenüber dem Computer und Selbsteinschätzungen der Sicherheit und Vertrautheit im Umgang mit Computeranwendungen erfassen, können der Klasse der Persönlichkeitstests zugeordnet werden.

2. Testtheorien und Testkonstruktion

Testtheorien sind in der Regel mathematisch formalisierte Theorien über den Zusammenhang zwischen dem beobachtbaren Testverhalten einer Person (z.B. Lösung einer Aufgabe, Zustimmung zu einer Aussage, gemessener Intelligenzquotient) und dem interessierenden Merkmal (z.B. Intelligenz, Einstellung zum politischen System der BRD). Generell wird implizit oder explizit davon ausgegangen, dass die Ausprägung des zu messenden Merkmals das Testverhalten beeinflusst – zumindest genau dann, wenn die Testitems das zu erfassende Merkmal auch *tatsächlich* erfassen, i.e. (konstrukt-)*valide* sind (s.o. Beitrag Dzeyk & Groeben). Gemeinsam ist weiterhin allen verfügbaren Testtheorien, dass die Konstruktion eines Tests auf ihrer Basis Möglichkeiten zur Verfügung

stellt, die *Reliabilität* eines Tests abzuschätzen und ggf. – falls diese als nicht zufriedenstellend angesehen werden muss – den Test zu revidieren.

2.1. Klassische Testtheorie und probabilistische (Item-Response-)Modelle

Wer die Literatur zur Testtheorie sichtet, wird eher früher als später auf die Unterscheidung zwischen ‚klassischer' und ‚probabilistischer' Testtheorie stoßen. Wir übernehmen in diesem Beitrag diese Einteilung aus pragmatischen Gründen, wiewohl neuere Überlegungen zeigen (vgl. Rost 1996; Steyer & Eid 1993), dass eine strikte Dichotomisierung zwischen ‚klassischen' und ‚probabilistischen' Modellen nicht aufrechterhalten werden kann. Da außerdem, wie Steyer und Eid (1993) u.E. zu Recht argumentieren, die Bezeichnung ‚probabilistische' Testtheorie missverständlich ist, weil der zentrale Unterschied zur klassischen Testthorie nicht erfasst wird, verwenden wir den Terminus ‚Item-Response-Theorie' (s. unten).

2.1.1. Klassische Testtheorie

Die klassische Testtheorie (KT) ist, wie der Name vermuten lässt, die historisch frühere der beiden Testtheorien (bzw. Klassen von Testtheorien); sie geht auf eine durch Gulliksen (1950) geleistete Systematisierung der bis dahin vorliegenden Versuche der Diagnose interindividueller Unterschiede zurück. Die zentralen Annahmen der KT sind die Folgenden:

1. Das zu messende Merkmal *i* hat bei der getesteten Person *v* eine definierte Ausprägung τ_{vi}. (z.B. Petra hat in Wahrheit einen Intelligenzquotienten von 115).
2. Der durch den Test gemessene Wert x_{vi} (z.B. ein bei Petra gemessener IQ von 110) setzt sich zusammen aus (a) der ‚wahren' Ausprägung des infragestehenden Merkmals (Petras IQ von 115) und (b) einem Fehlerwert ε_{vi} (Petra war bei der Durchführung des Intelligenztests müde und hat deswegen ein Testergebnis erreicht, das 5 Punkte unter ihrem ‚wahren' Wert liegt; der Fehlerwert betrüge folglich in diesem Beispiel -5). Umgekehrt wäre denkbar, dass bei Friedrich, dessen wahrer IQ 100 beträgt, ein IQ von 105 gemessen wird, da Friedrich bereits einige Aufgaben des Intelligenztests kannte. Der Fehler betrüge in diesem Falle +5.

   formal: $x_{vi} = \tau_{vi} + \varepsilon_{vi}$ (‚Verknüpfungsaxiom').

   Da diese Beziehung für *jede* Person *v* und für *jedes* Merkmal *i* Gültigkeit besitzen soll, kann auf die respektiven Indizes auch verzichtet werden und die Gleichung für die Variablen aller Messwerte, wahrer Werte und Fehlerwerte entsprechend formuliert werden:

   $$x = \tau + \varepsilon$$

3. Die Fehler treten lediglich zufällig und unsystematisch auf. Würde – hypothetisch – bei Petra der gleiche Intelligenztest unendlich häufig angewandt und der Durchschnitt der Ergebnisse jeder Testdurchführung gebildet, wäre das Ergebnis Petras wahrer Intelligenzquotient, da sich die Fehler bei den unterschiedlichen Testdurchführungen sozusagen gegenseitig ‚ausmitteln' würden.

   formal: $E(x_{vi}) = \tau_{vi}$ (‚Existenzaxiom')

   bzw. für jede Person und jedes Merkmal:

   $$E(x) = \tau$$

4. Aus den Annahmen (2) und (3) folgt, dass der Mittelwert der Fehlerwerte bei einer hypothetischen unendlich häufigen Wiederholung der Testung 0 betrüge.

   formal: $E(\varepsilon_{vi}) = 0$

   bzw. für jede Person und jedes Merkmal:

   $E(\varepsilon) = 0$

5. Aus der Annahme, dass die Fehlerwerte lediglich zufällig auftreten, lässt sich weiterhin ableiten, dass *keinerlei systematische gemeinsame Varianz* zwischen den Fehlerwerten zweier Personen beim gleichen Test oder zweier Tests bei der gleichen Person auftritt. Anschaulich: Wenn die Fehlerwerte in einer Reihe von Tests sowohl bei Petra als auch bei Friedrich zufällig verteilt sind, kann nicht vom Fehler, der für Petra in Test *i* auftritt, auf den Fehler, der für Friedrich in Test *i* auftritt, geschlossen werden. Umgekehrt kann, bearbeitet Petra zwei Tests, nicht von Petras Fehlerwert in Test *i* auf Petras Fehlerwert in Test *j* geschlossen werden. Weiterhin lässt sich folgern, dass keine systematische gemeinsame Varianz zwischen Fehlern und wahren Werten auftritt (beispielsweise dergestalt, dass die Testwerte von Personen mit höheren Merkmalsausprägungen in stärkerem Maße fehlerbehaftet wären).

Bei den Annahmen 2 und 3 handelt es sich um *Axiome*, also um solche Annahmen, die einer empirischen Überprüfung und damit auch einem Scheitern an der Wirklichkeit nicht zugänglich sind. Akzeptiert man jedoch diese Axiome (bzw. entschließt sich, einen Test auf ihrer Grundlage zu konstruieren), ergeben sich einige angenehme Konsequenzen: Aus den Annahmen 2 und 3 lässt sich ableiten, dass sich nicht nur der *Testwert* in wahren Wert und Fehler zerlegen lässt. Vielmehr gilt, dass sich auch die *Unterschiedlichkeit (Varianz) der Testwerte verschiedener Personen* (Petra und Friedrich) additiv aus der *Varianz der wahren Werte* und der *Varianz der Fehler* zusammensetzt. Auf dieser Basis lässt sich sodann eine sinnige Definition der Reliabilität eines Tests geben: Ein Test misst offensichtlich das, was er messen soll, umso reliabler (zuverlässiger), je eher sich Unterschiede in den Testwerten verschiedener Personen (IQ 110 bei Petra und IQ 105 bei Friedrich) darauf zurückführen lassen, dass die Personen *tatsächlich* unterschiedliche Merkmalsausprägungen besitzen (Petra: IQ 115 und Friedrich: IQ 100). Komplementär dazu sollten Unterschiede in den Testwerten *nicht* darauf zurückgehen, dass der Messfehler bei unterschiedlichen Personen (-5 bei Petra und +5 bei Friedrich) unterschiedlich ausgeprägt ist. Die Reliabilität eines Tests lässt sich also formal bestimmen als *Verhältnis der Varianz der wahren Werte zur Varianz der Testwerte.*

Da man nur die Testwerte, nicht jedoch die wahren Werte kennt, ist zunächst auch lediglich die Varianz der Testwerte bekannt. Die Reliabilität eines Tests kann dennoch auf die in Beitrag 1 (s.o. Dzeyk & Groeben) beschriebenen Arten bestimmt werden, die als Weg zur Schätzung des Verhältnisses der Varianz von wahren und Fehler-Werten jeweils aus den Axiomen der KT (insbesondere aus Annahme 5) ableitbar sind.

Für die auf der klassischen Testtheorie basierende Skala zur Erfassung einer Teilfähigkeit des Lesens (Geschwindigkeit lokaler Kohärenzbildung nach dem Modell von van Dijk und Kintsch 1983) lässt sich beispielsweise als Schätzung für die Reliabilität eine interne Konsistenz (Cronbach's α, s.o. Beitrag Dzeyk & Groeben) von .85 bestimmen: 85% der Unterschiede in den Testwerten wären hiernach darauf zurückzuführen, dass die unterschiedlichen Testpersonen beim Lesen eines Textes tatsächlich unterschiedlich

schnell lokale Kohärenz herstellen können. Komplementär dazu gehen 15% der Unterschiedlichkeit der Testwerte darauf zurück, dass für unterschiedliche Personen unterschiedliche Fehler auftreten.

2.1.2. Item-Response-Theorie

Während die KT zunächst lediglich die Existenz einer Merkmalsausprägung behauptet, im Weiteren auf den als Annahmen (2) und (3) eingeführten Axiomen aufruht und damit lediglich die Beziehung zwischen dem *Gesamttestwert* und der Merkmalsausprägung zum Gegenstand hat, machen Testmodelle auf Basis von Item-Response-Theorien („IRT-Modelle") weitergehende Annahmen, die mehrheitlich einer empirischen Prüfung zugänglich sind. Zentral ist dabei die modellhafte Spezifikation der Beziehung zwischen der beobachtbaren Reaktion auf *jedes einzelne Testitem* und der Ausprägung des zu messenden latenten Merkmals.

Die frühen IRT-Modelle (z.B. Rasch 1960; Birnbaum 1968) beziehen sich durchweg auf Items mit dichotomer Antwortskala (Lösung vs. Nichtlösung, Zustimmung vs. Ablehnung) und kontinuierliche latente Merkmale. Neuere Entwicklungen haben diese Beschränkung aufgehoben und auf der Basis des Modells von Rasch (1960) auch IRT-Modelle für mehrstufige bzw. kontinuierliche Items und kategoriale latente Merkmale wie Gruppenzugehörigkeiten formuliert (vgl. z.B. Müller 1987; Rost 1996). Der Fall dichotomer Antwortformate und ausschließlich kontinuierlicher latenter Variablen kann damit heute als „eher uninteressanter Spezialfall" (Rost 1996, 94) von solchen Modellen gelten, mit denen mehrstufige Antwortformate und sowohl kategoriale als auch kontinuierliche latente Merkmale behandelt werden können. Wir werden hier nichtsdestoweniger lediglich diesen „uninteressanten Spezialfall" abhandeln, weil er die am einfachsten zu handhabende Anwendung von IRT-Modellen darstellt (für die Generalisierungen auf mehrstufige oder kontinuierliche Items sowie auf kategoriale latente Variablen vgl. die angegebene Literatur).

Spezifiziert wird bei IRT-Modellen, bei welcher Ausprägung des zu messenden Merkmals welche Wahrscheinlichkeit für welche Bearbeitung eines Items (Lösung einer Aufgabe in einem Intelligenztest, Zustimmung zu einer Aussage in einem Einstellungstest) resultiert. Anders ausgedrückt: Es wird eine Funktion formuliert, die die Abhängigkeit der Lösungswahrscheinlichkeit der Items eines Tests von der Ausprägung des zu messenden Merkmals beschreibt (üblicherweise als „itemcharakteristische-" oder *IC-Funktion* bezeichnet).

Anschaulich: Gegeben sei wiederum eine Aufgabe aus einem Intelligenztest, die einen bestimmten Schwierigkeitsgrad aufweist. Für eine sehr intelligente Person A sollte nun eine hohe Wahrscheinlichkeit bestehen, dass sie die Aufgabe löst. Für eine weniger intelligente Person B sollte diese Wahrscheinlichkeit niedriger sein. Umgekehrt sollten die Personen A und B eine leichtere Aufgabe jeweils mit einer höheren Wahrscheinlichkeit lösen und eine schwierigere mit einer niedrigeren Wahrscheinlichkeit.

Diesen Sachverhalt illustriert Abbildung 1 für drei Items: Auf der Abszisse ist die Ausprägung des gemessenen Merkmals (ξ) abgetragen, auf der Ordinate die Lösungs- oder Zustimmungswahrscheinlichkeit. Die drei Kurven beschreiben die Lösungs- oder Zustimmungswahrscheinlichkeiten der Items in Abhängigkeit von der Ausprägung des gemessenen Merkmals. Wie aus der Abbildung ersichtlich ist, wird die jeweils gleiche Wahrscheinlichkeit, das Item zu lösen (z.B. 0.5) für Item 1 schon bei einer vergleichs-

weise niedrigen Merkmalsausprägung (mit β_1 bezeichnet) erreicht. Um Item 2 mit einer Wahrscheinlichkeit von 0.5 zu lösen, ist eine höhere Ausprägung des gemessenen Merkmals erforderlich (β_2) und für Item 3 die höchste (β_3). Man kann davon sprechen, dass Item 3 die höchste *Schwierigkeit* (β_3) aufweist (und Item 1 entsprechend die niedrigste (β_1)).

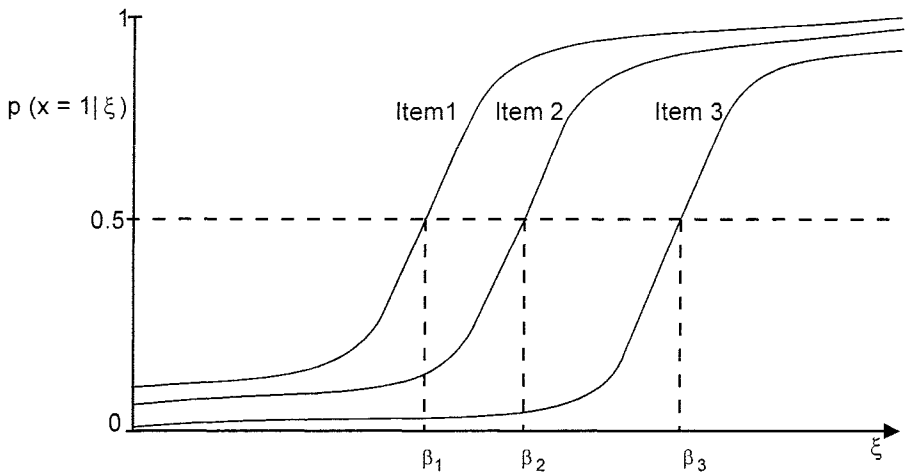

Abbildung 1: Verlauf der IC-Funktionen dreier Rasch-skalierter Items unterschiedlicher Schwierigkeit

Die in Abbildung 1 dargestellten Funktionsverläufe stellen lediglich eine mögliche Spezifizierung des Zusammenhangs von Testverhalten und Merkmalsausprägung dar, nämlich jene, die das sogenannte Rasch-Modell (Rasch 1960) vorsieht.[1] Ob die Eigenschaften der Items eines Tests den durch das jeweilige Testmodell spezifizierten Voraussetzungen genügen (z.B. dass alle Items des Tests eine IC-Funktion eines bestimmten Typs aufweisen), kann jeweils empirisch überprüft werden. Auf diese Weise wird es durch die Anwendung von IRT-Modellen möglich, die Items daraufhin zu beurteilen, ob sie *tatsächlich* dasselbe Merkmal erfassen oder nicht – also eine Überprüfung eines Aspekts der *Konstruktvalidität* des Tests (s.o. Beitrag Dzeyk & Groeben) vorzunehmen. Hiermit ist ein wichtiger Vorteil von IRT-Modellen gegenüber der klassischen Testtheorie markiert (die lediglich eine Theorie der *Reliabilität* von Tests bietet). Ein weiterer wesentlicher Vorteil der voraussetzungsreicheren IRT-Modelle gegenüber der KT besteht darin, dass die Beurteilung von zentralen Eigenschaften der Testitems (z.B. ihrer Schwierigkeit) unabhängig von der (kontingenten) Zusammensetzung einer Personenstichprobe, der der Test zu Konstruktionszwecken vorgelegt wurde, möglich wird.

1 In Abhängigkeit vom mathematischen Typ der IC-Funktion, der Anzahl der getroffenen Modellrestriktionen und dem Skalenniveau (s. Abschnitt 3.1.) des latenten Merkmals lassen sich unterschiedliche IRT-Modelle voneinander abgrenzen (vgl. z.B. Birnbaum 1968; Mokken 1971; Moosbrugger & Müller 1981; Haertel 1994).

Zusammenfassend lässt sich resümieren, dass IRT-Modelle gegenüber der KT die theoretisch besser explizierten und empirisch reichhaltigeren Testmodelle darstellen. Dieser Gewinn an Präzision wird allerdings um den Preis erkauft, dass es vergleichsweise schwierig ist, Itemmengen zu finden, die den Anforderungen eines IRT-Modells genügen. Nicht zuletzt hierin dürfte der Grund dafür liegen, dass nach wie vor die große Mehrzahl der in der psychologischen (und soziometrischen) Forschungs- und Diagnosepraxis zum Einsatz gebrachten Testverfahren auf Basis der KT konstruiert ist.

3. Probleme bei der Anwendung von Tests und Fragebogen

Die im Folgenden angesprochenen Probleme betreffen Test- und Fragebogenverfahren in unterschiedlichem Maße. Die Frage nach der messtheoretischen Fundierung stellt sich lediglich für psychologische Tests (Abschnitt 3.1.). Eine Berücksichtigung der psychologischen Prozesse, die dem Antwortverhalten zu Grunde liegen, ist dagegen sowohl für Tests wie für Fragebogen relevant (Abschnitt 3.2.). Dasselbe gilt für die sozialen Aspekte von Test- und Befragungssituation (Abschnitt 3.3.). Die zuletzt angesprochenen ethischen Probleme beziehen sich dagegen wieder in erster Linie auf psychologische Tests (Abschnitt 3.4.).

3.1. Zur messtheoretischen Fundierung psychologischer Testverfahren

Insofern mit psychologischen Testverfahren psychologische Merkmale gemessen werden sollen, sind die Begriffe ‚Messen' und ‚Testen' synonym. Gleichwohl entstammen beide Konzepte unterschiedlichen psychologischen Traditionen (vgl. Mausfeld 1983): Während der Testbegriff seine Wurzeln in der seit ihrem Beginn sozialtechnologisch orientierten Persönlichkeitsdiagnostik hat, wurde der Begriff des Messens zunächst vorwiegend im Kontext der Psychophysik gebraucht, wo es um eine exakte quantitative Beschreibung geht, in welcher Weise Empfindungseindrücke von physikalischen Größen abhängen. Aus der psychophysikalischen Tradition heraus hat sich die Messtheorie als eigenständiges Forschungsfeld verselbständigt, das eine mathematisch präzise Klärung der Voraussetzungen der Messbarkeit psychischer Merkmale zum Ziel hat (vgl. Orth 1974).

Auf Grund der historisch begründeten Trennung von Persönlichkeitsdiagnostik, Psychophysik und Messtheorie steht bei vielen psychologischen Testverfahren statt des grundlegenden Kriterium der Messbarkeit der zu erfassenden Merkmale das pragmatische Kriteriums der Bewährung psychologischer Tests für sozialtechnologische Zwecke im Vordergrund. Gleichwohl beruht der Rückschluss von Testwerten auf die Ausprägung eines psychologischen Merkmals immer auf bestimmten, meist impliziten theoretischen Annahmen.

Um welche Annahmen kann es sich dabei handeln? Aus messtheoretischer Sicht ist eine Messung eine Zuordnung von Zahlen zu Objekten oder Ereignissen (z.B. Testpersonen oder der Beurteilung eines Items durch eine Testperson), die die Bedingung einer homomorphen Abbildung eines empirischen auf ein numerisches Relativ erfüllt (Orth 1983). Unter einer homomorphen Abbildung ist eine Abbildung zu verstehen, bei der die Struktur der Relationen zwischen den Elementen des empirischen Relativs im nume-

rischen Relativ erhalten bleiben. Die Skala TECOWI aus dem Inventar zur Computerbildung (INCOBI), die das Merkmal ‚deklaratives Computerwissen' bei Studierenden erfassen soll, beruht z.B. auf der Voraussetzung, dass zwischen den Testpersonen der Zielpopulation tatsächlich bestimmte Relationen hinsichtlich dieses Merkmals bestehen (z.B. „Der Unterschied zwischen einer computererfahrenen Person A und einer weniger computererfahrenen Person B ist genauso groß wie der Unterschied zwischen den entsprechenden Personen C und D."), und dass sich diese Relationen in den Testwerten widerspiegeln („Die Testwertdifferenz zwischen Person A und Person B ist genauso groß wie der Unterschied zwischen Person C und Person D."). Die Frage der Existenz einer Skala in diesem Sinne ist Gegenstand des sog. *Repräsentationsproblems*. Weitere grundlegende messtheoretische Probleme sind das *Eindeutigkeits-* und das *Bedeutsamkeitsproblem*. Das Eindeutigkeitsproblem betrifft die Transformationen, denen die Abbildungsfunktion unterworfen werden kann, ohne dass sich die Skaleneigenschaften ändern. Die Eigenschaften der Skala TECOWI sollten sich beispielsweise nicht ändern, wenn man zu den Testwerten der einzelnen Personen eine konstante Zahl addiert. Bei dem Bedeutsamkeitsproblem geht es um die Frage, welche mathematischen Operationen auf die erhobenen Testwerte sinnvoll angewandt werden können. In unserem Beispiel haben wir die Skala TECOWI so beschrieben, dass sie die Voraussetzungen einer Intervallskala erfüllen soll. Für Messungen auf diesem Skalenniveau ist die Berechnung von arithmetischem Mittel, Varianzen und Produkt-Moment-Korrelationen sinnvoll, was entscheidend für die Anwendung wichtiger statistischer Verfahren ist. Für Messungen auf Nominalskalenniveau, durch die Objekte kategorialen Merkmalen (z.B. Geschlecht oder Studienfach) zugeordnet werden, oder Messungen auf Ordinalskalenniveau, bei denen die Elemente des numerischen Relativs lediglich Ordnungsrelationen (z.B. „ist größer als") abbilden, gelten stärkere Beschränkungen hinsichtlich der Bedeutsamkeit mathematischer Operationen. Deshalb wird mit den meisten psychologischen Tests eine Messung auf Intervallskalenniveau angestrebt.

Die messtheoretische Spezifikation der Beziehung zwischen einem latenten psychologischen Merkmal und den beobachtbaren Messwerten trägt jedoch nur wenig zu einem psychologischen Verständnis des tatsächlichen Verhaltens der Befragten in der Testsituation bei. Hierzu müssen die psychologischen Prozesse herangezogen werden, die der Testbearbeitung zugrundeliegen.

3.2. Psychologische Prozesse bei der Bearbeitung von Fragebogen und Tests

Das Grundprinzip des psychologischen Testens (vgl. Abschnitt 1) stammt aus dem methodologischen Behaviorismus, insofern die Testitems als standardisierte Reize verstanden und zur Ermittlung des Testwerts lediglich die beobachtbaren Reaktionen der Testpersonen herangezogen werden. Wie diese Reaktionen im individuellen Fall zustandekommen, ist aus psychometrischer Sicht vollkommen uninteressant. Trotzdem sind auch psychometrische Annahmen psychologisch gehaltvoll. So wird beispielsweise in der Regel von einer Testbearbeitung ausgegangen, bei der die Beantwortung eines Items von der Beantwortung vorheriger Items unabhängig ist. Eine entsprechende Voraussetzung für die Berechnung von Retest-Reliabilitäten (s.o. Beitrag Dzeyk & Groeben) besteht

darin, dass die Bearbeitung des Tests zum zweiten Messzeitpunkt von der erstmaligen Vorgabe nicht beeinflusst wird.

Empirisch lässt sich jedoch zeigen, dass psychometrische Voraussetzungen dieser Art häufig nicht erfüllt sind. Insbesondere in der Einstellungs- und Umfrageforschung existiert umfangreiche Literatur zu *Kontexteffekten* (für einen Überblick vgl. z.B. Tanur 1992; Schwarz & Sudman 1992; zu Persönlichkeitstests vgl. z.B. Knowles 1988). Ein Kontexteffekt kann z.B. darin bestehen, dass in einem Fragebogen zur Einstellung gegenüber der Legalität von Abtreibungen ein Item, das die freie Entscheidung der Frau thematisiert („Soll eine Abtreibung legal sein, wenn eine Frau verheiratet ist und keine Kinder mehr haben will?"), dann weniger häufig positiv beantwortet wird, wenn es nach einem Item präsentiert wird, das sich auf den Fall einer pränatalen Schädigung des Kindes bezieht („Soll eine Abtreibung legal sein, wenn die Wahrscheinlichkeit einer schwerwiegenden Schädigung des Babies hoch ist?"; vgl. Schuman & Presser 1981). Andere Typen von Kontexteffekten treten in Abhängigkeit von der Frageformulierung, der Benennung der Antwortkategorien oder der Reihung von allgemeineren und spezielleren Fragen auf (eine Systematik bieten Tourangeau & Rasinski 1988).

Kontexteffekte zeigen, dass Personen, die einen Einstellungs- oder Persönlichkeitstest bearbeiten, auf die Items nicht qua objektiv gegebene Stimuli reagieren. Vielmehr stellt die Itembeantwortung eine komplexe kognitive Leistung dar, die auf dem Zusammenwirken einer Vielzahl mentaler Prozesse beruht: Eine Person interpretiert die Testitems im Hinblick auf ihr Verständnis der Testsituation, beurteilt sie vor dem Hintergrund ihrer Überzeugungen über die fragliche Einstellungsthematik bzw. die eigene Person und muss schließlich ihr Urteil mit dem Antwortformat abstimmen, das ihr durch den Test vorgegeben wird (vgl. das kognitive Modell der Itembeantwortung bei Tourangeau & Rasinksi 1988). Ein gravierendes Validitätsproblem liegt vor, wenn die Items eines Einstellungsfragebogens oder Persönlichkeitstests Sachverhalte ansprechen, die eine Person nicht zuverlässig beurteilen kann, etwa weil sie noch nie darüber nachgedacht hat (zum sog. Nonattitudes-Problem vgl. Converse 1970). Fragen nach objektiven Sachverhalten, z.B. nach der Häufigkeit eines bestimmten Verhaltens, können verzerrenden Gedächtniseffekten unterliegen, die teilweise durch eine geeignete Frageformulierung und -reihung zu vermeiden sind (vgl. z.B. Strube 1987). Bei der Konstruktion der im Projekt entwickelten Skalen, die auf Selbstauskünften und Selbsteinschätzungen beruhen, wurde versucht, die der Beantwortung zu Grunde liegenden kognitiven Prozesse so weit wie möglich im Blick zu haben. Beispielsweise sind die Teilskalen des Inventars zur Computerbildung (INCOBI) so angeordnet, dass die für Kontexteffekte anfälligeren Skalen zur Erfassung computerbezogener Einstellungen am Anfang, die robusteren Wissenstests hingegen am Ende bearbeitet werden. Dem Nonattitudes-Problem begegnen wir, indem den Respondenten/innen die Möglichkeit eingeräumt wird, einzelne Items oder ganze Skalen nicht zu bearbeiten, wenn sie sich zu einem Urteil nicht in der Lage sehen. Die Fragebogen, mit denen objektive Daten zur tatsächlichen Computernutzung (oder zu Lese- und Schreibaktivitäten) erhoben werden, sind so strukturiert, dass die Erinnerungsleistungen der Probanden/innen gezielt gestützt werden.

Die Erforschung der kognitiven Grundlagen des Antwortverhaltens beschränkt sich nicht auf Fragebogen, Persönlichkeits- und Einstellungstests. Für den Bereich der Leistungsmessung existieren analoge Versuche, kognitive Modelle der Aufgabenbearbeitung

zu formulieren. Als ein Beispiel sei hier der Kognitive-Komponenten-Ansatz der Intelligenzforschung genannt, in dessen Rahmen die Bearbeitung verschiedener Aufgabentypen im Hinblick auf das Zusammenwirken kognitionspsychologisch differenzierbarer Teilprozesse untersucht wird (vgl. Waldmann 1996).

3.3. Testung und Befragung als soziale Situationen

Die Bearbeitung eines Tests oder eines Fragebogens fordert von den Respondenten/innen nicht nur eine komplexe kognitive Leistung, sondern auch die Teilnahme an einer sozialen Interaktion mit den Forschenden. Insbesondere Persönlichkeits- und Einstellungstests, aber auch Fragebogen unterliegen daher besonderen Einflüssen durch soziale Wahrnehmungen und Ziele der Befragten. In die methodische Diskussion hat dieser Umstand vor allem unter den Schlagworten *Selbstdarstellung* und *soziale Erwünschtheit* Eingang gefunden (vgl. Mummendey 1987).

Mit Selbstdarstellung ist gemeint, dass Respondenten/innen die Testsituation als kommunikative Situation auffassen, in der sie dem/der Forscher/in (und gegebenenfalls auch sich selbst) etwas über die eigene Person, über ihr eigenes Selbstkonzept oder ihre eigenen Überzeugungen mitteilen. Die Antworten können demnach in Abhängigkeit von dem Sinn, den die Respondenten/innen der Befragung beilegen, variieren (Mummendey 1990). Die soziale Erwünschtheit stellt einen Spezialfall der Selbstdarstellung dar, bei dem die Test- oder Fragebogenitems generell mit dem Ziel beantwortet werden, ein möglichst positives Bild der eigenen Person im Hinblick auf soziale Normen und Standards zu präsentieren.

Während Selbstdarstellung im Allgemeinen ein ubiquitäres Phänomen in Test- und Befragungssituationen sein dürfte, tritt das spezielle Problem einer Antwortverzerrung durch soziale Erwünschtheit vor allem in bestimmten individualdiagnostischen Zusammenhängen, etwa bei der Selektion von Stellenbewerbern/innen auf. Für Testanwendungen in der Forschung empfiehlt sich u.E. weniger die statistische Kontrolle mit Hilfe gesonderter Erwünschtheits- oder ‚Lügen'-Skalen (vgl. z.B. Edwards 1970) als vielmehr eine konsequente Berücksichtigung des ‚Testnebengütekriteriums' der Transparenz. Aus der Perspektive eines epistemologischen Subjektmodells (vgl. Groeben, Wahl, Schlee & Scheele 1988) sind valide Selbstauskünfte vor allem dann zu erwarten, wenn für die Befragten Sinn und Zweck der Untersuchung sowie die Messintention der angewendeten Skalen in möglichst hohem Maße durchschaubar sind. Entsprechend stellt eine transparente Testgestaltung ein wesentliches Konstruktionsprinzip aller im Rahmen unseres Projekts entwickelten Instrumente dar. Schließlich sollte die ausdrückliche Zusicherung einer Anonymisierung der erhobenen Daten selbstverständlich sein, insbesondere da viele Probanden/innen psychologische Tests fälschlicherweise ausschließlich mit einem individualdiagnostischen Erkenntnisinteresse in Verbindung bringen (vgl. Bortz & Döring 1995, 215).

3.4. Ethische Probleme der Testanwendung

Neben einer möglichst transparenten Testgestaltung erfordert eine verantwortliche Anwendung von Tests die Berücksichtigung weiterer ethisch relevanter Aspekte, von denen

hier nur zwei, für die Testanwendung spezifische Probleme erwähnt werden sollen (weitere Hinweise bei Haecker, Leutner & Amelang 1998; Messick 1982).

Für viele Leistungstests, Persönlichkeitstests und Einstellungsfragebogen gilt, dass die Testbearbeitung für die Testperson auch eine Konfrontation mit den eigenen Fähigkeiten, dem eigenen Selbstbild oder den eigenen Überzeugungen über möglicherweise heikle Themen bedeutet. Dieser Umstand kann in individuell unterschiedlichem Maße zu *psychischen Belastungen* und Frustrationen führen, auf die seitens der Testanwender/innen angemessen zu reagieren ist.

Ein weiteres Problem betrifft die *Testfairness*, die vor allem für Auslesesituationen diskutiert worden ist, aber auch für die Interpretation sozialwissenschaftlicher Untersuchungen bedeutsam sein kann (vgl. z.B. Jensen 1980). Die Testfairness bemisst sich danach, in welchem Ausmaß ein Test bestimmte soziale Gruppen (z.B. ethnische Gruppierungen, soziale Schichten, Absolventen/innen verschiedener Schulformen etc.) systematisch benachteiligt, d.h. zu konsistent schlechteren oder sozial negativer bewerteten Ergebnissen in diesen Gruppen führt. Die Frage, inwieweit die Fairness eines Tests schon bei seiner Konstruktion berücksichtigt werden sollte, ist äußerst diffizil. Für allgemeine Leistungstests, wie etwa Intelligenztests, kann man z.B. begründeterweise fordern, dass die Testwerte sich nicht systematisch zwischen Personen verschiedenen Geschlechts unterscheiden sollten. Wenn allerdings hinsichtlich eines Merkmals tatsächliche Differenzen zwischen bestimmten sozialen Gruppen zu erwarten sind, würde die Anwendung entsprechender Fairnesskriterien bei der Testkonstruktion zu einer artifiziellen Verschleierung dieser Unterschiede führen. Dies betrifft etwa die Computer Literacy-Skalen des Inventars zur Computerbildung (INCOBI), die auch sensitiv gegenüber eventuellen Unterschieden zwischen weiblichen und männlichen Studierenden sein sollen.

4. Fazit

Psychologische Tests und Fragebogen, die sich per Definition durch einen hohen Grad von Standardisierung und Objektivität auszeichnen, können eine sinnvolle und ökonomische Methode der sozialwissenschaftlichen Datengewinnung sein. Die Erfüllung der Gütekriterien Reliabilität und Validität hängt dabei nicht nur davon ab, inwieweit bei der Testkonstruktion psychometrische (test- und messtheoretische) Gesichtspunkte herangezogen wurden. Die Testbearbeitung stellt eine komplexe kognitive Leistung dar, die im Rahmen einer sozialen Situation erbracht wird. Dies muss bei der Testkonstruktion, -anwendung und -interpretation in Rechnung gestellt werden. Darüber hinaus erfordert die Erfüllung von Validitätskriterien in vielen Fällen eine aussagekräftige Theorie, auf der die Testkonstruktion basiert. Dazu gehört an zentraler Stelle eine dezidierte Vorstellung darüber, welche Indikatoren für die Erfassung des interessierenden psychologischen Merkmals besonders geeignet sind. Qualitative Methoden können in diesem Zusammenhang zu einer vorstrukturierenden Erfassung des angezielten Gegenstandsbereichs eingesetzt werden (vgl. z.B. die Durchführung halbstrukturierter Interviews zur Entwicklung der Skalenkonzeption und der Zusammenstellung von Testitems).

5. Literatur

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examenee's ability. In F. M. Lord & M. R. Nowick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation für Sozialwissenschaftler* (2. Aufl.). Berlin: Springer.

Brickenkamp, R. (1996). *Handbuch psychologischer und pädagogischer Tests* (2. Aufl.). Göttingen: Hogrefe.

Christmann, U., Groeben, N., Flender, J., Naumann, J. & Richter, T. (1999). Verarbeitungsstrategien von traditionellen (linearen) Buchtexten und zukünftigen (nichtlinearen) Hypertexten. In N. Groeben (Hrsg.), *Lesesozialisation in der Mediengesellschaft. Ein Schwerpunktprogramm (10. Sonderheft IASL)* (S. 175–189). Tübingen: Niemeyer.

Converse, P. (1970). Attitudes and nonattitudes: Continuation of a dialogue. In E. Tufte (Ed.), *The quantitative analysis of social problems* (pp. 168–189). Reading, MA: Addison-Wesley.

Edwards, A. L. (1970). *The measurement of personality traits by scales and inventories.* New York, NY: Holt, Rinehart and Winston.

Eysenck, H. J. (1953). Fragebogen als Meßmittel der Persönlichkeit. *Zeitschrift für Experimentelle und Angewandte Psychologie, 1,* 291–335.

Groeben, N., Wahl, D., Schlee, J. & Scheele, B. (1988). *Das Forschungsprogramm Subjektive Theorien. Eine Einführung in die Psychologie des reflexiven Subjekts.* Tübingen: Francke.

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

Haecker, H., Leutner, D. & Amelang, M. (1998). Standards für pädagogisches und psychologisches Testen. *Diagnostica, Suppl. 1.*

Haertel, E. H. (1994). Continuous and discrete latent structure models for item response data. *Psychometrika, 55,* 477–494.

Jensen, S. R. (1980). *Bias in mental testing.* London: Methuen.

Knowles, E. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55,* 312–320.

Lienert, G. A. & Raatz, U. (1994). *Testaufbau und Testanalyse* (5. Aufl.). Weinheim: Beltz.

Mausfeld, R. (1994). Von Zahlzeichen zu Skalen. In W. H. Tack & T. Herrmann (Hrsg.), *Methodologische Grundlagen der Psychologie. Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie 1: Forschungsmethoden der Psychologie, Bd. 1* (S. 565–603). Göttingen: Hogrefe.

Messick, S. (1982). Test validity and the ethics of assessment. *Diagnostica, 28,* 1–25.

Mokken, H. (1971). *A theory and procedure of scale analysis.* Den Haag: Mouton.

Moosbrugger, H. & Müller, H. (1981). Ein klassisches latent-additives Testmodell (KLA-Modell). In W. Michaelis (Hrsg.), *Bericht über den 32. Kongreß der Deut-*

schen Gesellschaft für Psychologie in Zürich 1980, Band 2 (S. 483–486). Göttingen: Hogrefe.

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika, 52*, 165–181.

Mummendey, H. D. (1987). *Die Fragebogen-Methode*. Göttingen: Hogrefe.

Mummendey, H. D. (1990). *Psychologie der Selbstdarstellung*. Göttingen: Hogrefe.

Orth, B. (1974). *Einführung in die Theorie des Messens*. Stuttgart: Kohlhammer.

Orth, B. (1983). Grundlagen des Messens. In H. Feger & J. Bredenkamp (Hrsg.), *Messen und Testen. Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie 1: Forschungsmethoden der Psychologie, Bd. 3* (S. 136–180). Göttingen: Hogrefe.

Rasch, G. (1960). *Studies in mathematical psychology I: Probabilistic models for some intelligence and attainment tests*. Kopenhagen: Nielsson & Lydiche.

Richter, T., Naumann, J. & Groeben, N. (im Druck). Das Inventar zur Computerbildung (INCOBI): Ein Instrument zur Erfassung von Computer Literacy und computerbezogenen Einstellungen bei Studierenden der Geistes- und Sozialwissenschaften. *Psychologie in Erziehung und Unterricht*.

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

Schuman, J. & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments in question form, wording, and context*. New York: Academic Press.

Schwarz, N. & Sudman, S. (Eds.) (1992). *Context effects in social and psychological research*. Berlin: Springer.

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

Strube, G. (1987). Answering survey questions: The role of memory. In H. J. Hippler, N. Schwarz & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 86–101). Berlin: Springer.

Tanur, J. M. (Ed.) (1992). *Questions about questions – inquiries into the cognitive bases of surveys*. New York, NY: Sage.

Tourangeau, R. (1992). Context effects on responses to attitude questions: Attitudes as memory structures. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 35–48). Berlin: Springer.

Tourangeau, R. & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*, 299–314.

van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

Waldmann, M. R. (1996). Kognitionspsychologische Theorien von Begabung und Expertise. In F. E. Weinert (Hrsg.), *Psychologie des Lernens und der Instruktion. Enzyklopädie der Psychologie, Themenbereich D: Praxisgebiete, Serie I: Pädagogische Psychologie, Bd. 2* (S. 445–476). Göttingen: Hogrefe.

Anschrift der Autoren / der Autorin:
Johannes Naumann, Tobias Richter, Psychologisches Institut, Lehrstuhl II
Universität zu Köln, Herbert-Lewin-Str. 2, D-50931 Köln
E-Mail: johannes.naumann@uni-koeln.de
E-Mail: tobias.richter@uni-koeln.de
Ursula Christmann, Psychologisches Institut,
Universität Heidelberg, Hauptstr. 47–51, D-69117 Heidelberg
E-Mail: ursula_christmann@psi-sv2.psi.uni-heidelberg.de