Zitierhinweis:

Legree, P. J., Psotka, J., Tremble, T. & Bourne, D. R. (2006). Die Verwendung konsensbasierter Messverfahren zur Erfassung emotionaler Intelligenz. In R. Schulze, P. A. Freund & R. D. Roberts (Hrsg.), *Emotionale Intelligenz. Ein internationales Handbuch* (S. 165-189). Göttingen: Hogrefe.

8

Die Verwendung konsensbasierter Messverfahren zur Erfassung emotionaler Intelligenz

Peter J. Legree
Joseph Psotka
Trueman Tremble
U.S. Army Research Institute for the Behavioral and Social Sciences, USA
Dennis R. Bourne
American Psychological Association, USA

Vor mehr als einem Jahrhundert schrieb der russische Schriftsteller Leo N. Tolstoi: „Glückliche Familien sind alle gleich; jede unglückliche Familie ist auf ihre eigene Weise unglücklich."

Zusammenfassung

In den Bereichen der Arbeits- und Organisationspsychologie sowie der kognitiven Psychologie wurden Situational Judgment Tests (SJTs) entwickelt, um bestimmte Leistungen vorherzusagen und Kognitionstheorien bewerten zu können. Die Erstellung solcher Skalen erfordert für gewöhnlich die Meinungen und Ansichten von Experten auf dem jeweiligen Gebiet, um Auswertungsschlüssel oder Kriteriumsdaten zur Berechnung empirischer Standards zu erhalten. Eine einfachere und elegantere Prozedur wird in diesem Kapitel vorgestellt. Sie erlaubt, die Antworten der Probanden als Abweichungen von einem Konsens zu werten, der zuvor durch die Antwortverteilungen in der Probandenstichprobe definiert wurde. Dieser Ansatz wird *konsensbasierte Messung* (KBM) oder *Scoring* genannt und ist bereits zur Validierung von Skalen in Bereichen wie beispielsweise der emotionalen Intelligenz verwendet worden. In diesen Bereichen gibt es zwar ausgewiesene Experten,

aber kein klar abgegrenztes, objektives Wissen. Die zu diesem Paradigma vorliegenden Daten werden in diesem Kapitel zusammenfassend dargestellt; es finden sich sehr gute Übereinstimmungen zwischen den in SJTs erzielten Werten, für die die Auswertungsstandards einerseits auf Experten- und andererseits auf Probandenantworten basieren und für die entsprechend große Datenmengen vorliegen. Die Übereinstimmungen weisen darauf hin, dass Probandenantworten für das Scoring von SJTs verwendet werden können, wenn Expertenmeinungen nicht vorhanden sind. Validitätsdaten zu Skalen situationaler Beurteilung, die mit diesem Ansatz ausgewertet wurden, werden ebenfalls zusammenfassend beschrieben.

8.1 Einleitung

Im Laufe des letzten Jahrzehnts wurden szenariobasierte Tests zur Messung von Wissen und Expertise in Leistungsbereichen wie zum Beispiel Führung und Fahrsicherheit entwickelt, ebenso wie zur Erfassung emotionaler, sozialer und allgemeiner Intelligenz (Legree, 1995; Legree, Heffner, Psotka, Martin & Medsker, 2003; Legree, Martin & Psotka, 2000; Mayer, Salovey, Caruso und Sitarenios, 2003; McDaniel, Morgeson, Finnegan, Campion & Braverman, 2001).

Während die meisten Autoren dieser Anwendungen zur Entwicklung der Auswertungsschlüssel Expertengruppen beauftragten (siehe Hedlund et al., 2003), wurden für andere Verfahren Auswertungsschlüssel konstruiert, die auf Daten basierten, welche an großen Teilnehmergruppen erhoben wurden, die zwar auf dem entsprechenden Themengebiet kenntnisreich waren, aber nicht als Experten bezeichnet werden konnten (Legree, 1995; Legree et al., 2000, 2003). Die Auswertungsschlüssel dieser Gruppen von Nicht-Experten näherten sich stark den Expertenstandards an. In diesen früheren Artikeln wurde der Gebrauch von Nicht-Expertengruppen zur Entwicklung von Auswertungsstandards *Konsens-Scoring*, oder allgemeiner *konsensbasierte Messung* (KBM) genannt. KBM stellt eine auf maximaler Leistung basierende Methode zur Erfassung wissensrelevanter Konstrukte dar und ist für all jene Konzeptualisierungen emotionaler Intelligenz relevant, die Zusammenhänge von Wissen, Fertigkeiten und Fähigkeiten annehmen (siehe Kapitel 2 von Neubauer & Freudenthaler).

Die mit KBM verknüpfte Erwartung ist, dass das in der psychologischen Forschung untersuchbare Wissensspektrum so erweitert wird, dass Bereiche einbezogen werden können, für die keine *echten* Experten identifizierbar sind und in denen es kein objektives Faktenwissen gibt. KBM ist für die Messung emotionaler Intelligenz von Bedeutung, weil diese genau einem solchen Bereich zugeordnet werden kann. Es fehlen also noch immer genügend verfügbare Experten und objektives Wissen zu EI. Tatsächlich wird die theoretische Entwicklung auch noch weitgehend einem Stadium formativer Entwicklung zugeordnet. Ungeachtet dieser Tatsache wurde der Ansatz gewählt, um gut entwickelte leistungsbasierte Tests emotionaler Intelligenz auszuwerten, zu denen die Multi-factor Emotional Intelligence Scale (MEIS; Mayer, Caruso & Salovey, 1999) und der Mayer Salovey Caruso Emotional Intelligence Test (MSCEIT; Mayer et al., 2003) zählen. Die Vorstellung, dass Nicht-Experten zur Entwicklung des für das Antwortenscoring notwendigen „Experten"-Wissens eingesetzt werden, könnte jedoch bei Testentwicklern, die mit den Stärken und Schwächen dieses Ansatzes nicht vertraut sind, auf Ablehnung stoßen. Auch Kritiker haben diese Vorgehensweise in Frage gestellt (z. B. Roberts, Zeidner & Matthews, 2001; Schaie, 2001; Zeidner, Matthews & Roberts, 2001). Daher soll dieser Beitrag, in dem KBM und ihre Entwicklung in verschiedenen Bereichen der angewandten Psychologie genauer beschrieben werden (und in dem dazu relevante Daten

und Theorien zusammengefasst werden), dabei helfen, die Sachlage zu klären. Dazu werden wir ein Beispiel für die Verwendung dieses Ansatzes in unscharf eingegrenzten Wissensbereichen – wie der emotionalen Intelligenz – und in anderen Bereichen, in denen Experten möglicherweise nicht verfügbar sind, präsentieren.

8.2 Testkonstruktion in unscharf eingegrenzten Wissensbereichen

Viele psychologische Wissenstests basieren auf einer Arbeits- bzw. Aufgabenanalyse (oder einer kognitiven Analyse), die Wissens- und Leistungsbereiche miteinander verbindet. Dieser auf empirisch erhobenen Daten aufbauende Ansatz hat seinen Nutzen in vielen praktischen Bereichen nachgewiesen (siehe Anastasi & Urbina, 1997). Implizit sind hier die Erwartungen enthalten, dass formales und unterschwellig vorhandenes (implizites) Wissen die Grundlage für Leistung darstellt und Schlussfolgerungen für solches Wissen aus beobachtetem Verhalten anderer gezogen wird. Die Konstruktion von Wissenstests hat sich traditionell entweder auf verfügbares formales Wissen (wie von Experten geschriebene Bücher oder über Jahrzehnte durch Instruktion und Analyse weiterentwickelte pädagogische Materialien) oder auf die Ratschläge von angesehenen Experten gestützt.

Aber dennoch bleibt viel Wissen intuitiv und unterschwellig. Es könnte als reine Ansichtssache bezeichnet werden, und bisweilen gibt es für dieses Wissen keine formalen Wissensquellen bzw. nicht einmal Experten, die in der Lage sind, angemessene Standards zu definieren. Experten können oder scheinen in vielen Bereichen, wie Kunst, Musik, Politik, Regierung und Wirtschaftswissenschaften, deutlich andere Ansichten, Gedankengänge und Informationsquellen zu haben als die Vertreter jener Bevölkerungsschichten, die für Forscher im allgemeinen von Interesse sind. KBM stellt in solchen Situationen einen sinnvollen Ausweg aus diesem Dilemma dar.

8.2.1 Einschränkungen traditioneller Skalenkonstruktion

Obwohl KBM einigen Einschränkungen unterliegt, wollen wir darauf hinweisen, dass die traditionelle, auf Experten zurückgreifende Itemkonstruktion ebenfalls ihre Grenzen hat. Itemkonstruktion in formalen, klar definierten Wissensbereichen kann leicht allgemeines Wissen und Expertise reflektieren. Die anschließende Itemrevision beruht häufig auf der Verwendung von Itemstatistiken oder faktoranalytischen Techniken zur Maximierung von Skalencharakteristika wie Reliabilität und Validität. Weil Prädiktor- und Kriteriumsreliabilität die Skalenvalidität einschränken, ist die Maximierung der Testreliabilität von entscheidender Bedeutung. Testkonstruktionsentscheidungen haben daher häufig die Verbesserung der Skalenreliabilität zum Ziel. Zur Reliabilitätsmaximierung werden zumeist Itemstatistiken und besonders niedrige Itemtrennschärfen benutzt, um „schlecht“ abschneidende Items zu identifizieren. Aus der Perspektive der Item Response Theory resultieren diese Bemühungen darin, dass Items als ineffizient bei der Informationsgewinnung charakterisiert werden und verändert werden sollten. Testskalen werden durch die Auswahl einer Teilmenge von Items erzeugt, die zur Trennung von hoch- und niedrigleistenden Probanden geeignet sind. Die resultierenden Tests sind in der Regel reliabel und häufig hinsichtlich eines bestimmten Kriteriums valide.

Für viele akademische und wirtschaftliche Zwecke war dieser traditionelle Ansatz zur Entwicklung von Wissensmaßen, die für Personalmanagement und Trainingsentscheidungen sowohl valide als auch nützlich sind, ausreichend. Zum Beispiel ist ma-

thematisches Wissen gut untersucht und mit Leistungskriterien verknüpft, so dass es relativ einfach ist, die richtigen Antworten auf eine Vielzahl von Fragen zu finden, die das Verständnis grundlegender mathematischer Konzepte erfordern. Ebenso haben Wörter und Ausdrücke spezifische Bedeutungen und Nebenbedeutungen. Diese finden sich in Wörterbüchern. Wortschatzwissen wird häufig mit Aufgaben erfasst, denen diese Wörterbuchdefinitionen zu Grunde liegen. Diese Aufgaben werden auch häufig zur Messung allgemeiner Intelligenz eingesetzt. Die Aufgabenkonstruktion ist größtenteils aufgrund des Vorhandenseins von Expertenwissen möglich, welches für gewöhnlich die Verfügbarkeit von objektiven Informationen und gelegentlich die Meinungen der Experten widerspiegelt. Selbst einfache arithmetische und algebraische Probleme erfordern Expertise, obwohl diese weithin verfügbar ist. Die Bedeutung von Itemstatistiken liegt in der Erzeugung von Konsistenz innerhalb eines Messinstruments, so dass sehr gute Leistungen eines Probanden mit einer erhöhten Wahrscheinlichkeit, ein Item richtig zu beantworten, assoziiert werden. Aus der Perspektive der KBM erzeugt diese Prozedur in Wirklichkeit einen Konsens innerhalb der Standardisierungsgruppe. So gesehen werden also alle Skalen auf der Basis von Konsensentscheidungen konstruiert – KBM kann als eine Variante eines altbekannten Sachverhalts angesehen werden.

8.2.2 Wenn der Konsens nicht die beste Lösung ist

Offensichtlich werden gelegentlich Items generiert, für die Konsensmeinungen nicht richtig sind oder von denen verschiedene Gruppen von Menschen deutlich unterschiedliche Auffassungen haben: Was ist die Hauptstadt von Israel (Tel Aviv/Jerusalem)? Wo befindet sich die EU-Regierung (Brüssel/Straßburg)? Ist es richtig, dass die USA den Irak invadierten (Ja/Nein)? All dies sind Items, über die verschiedene Gruppen unterschiedliche Auffassungen haben oder für die verschiedene Auffassungen unterschiedliche Gültigkeit besitzen. Eine vernünftige Antwort auf solche gelegentlichen Meinungsverschiedenheiten ist nicht, die KBM zu verwerfen, sondern die Grundlage dieser Meinungsverschiedenheiten zu verstehen und dadurch Implikationen zu identifizieren, welche mit der Entwicklung und der Erfassung von Wissen und Meinungen verbunden sind. Des Weiteren ist die Möglichkeit interessant, dass das Wissen, das vielen Fragen zugrunde liegt, durch Analyse der Meinungen einer großen Anzahl von Nicht-Experten abgeleitet werden könnte.

8.2.3 Wissensbereiche, in denen es keine Experten gibt

Es ist unbestreitbar, dass es Wissensbereiche ohne Expertenwissen geben kann. Beispielsweise wäre vor den Anstrengungen von Noah Webster (1758–1843) die Erfassung des damaligen englischen Wortschatzes wohl relativ problematisch gewesen: Da eine geeignete Informationsquelle für Wortwissen (d. h. ein Wörterbuch) fehlte, wäre einem amerikanischen Vokabeltestentwickler des 18. Jahrhunderts nur die Möglichkeit geblieben, akzeptable Definitionen für amerikanische Ausdrücke wie „hickory“ und so gebräuchliche Ausdrücke wie „Bett“ über Expertenmeinungen zu bestimmen. Ob Experten die Anspielung auf einen Blumengarten als eine akzeptable Definiton für „Bett“ ansehen würden, bleibt eine offene Frage, aber die Richtung der Antwort würde die Auswertung des Tests beeinflussen.

Aber welche Bevölkerungsschicht hätte die geeigneten inhaltlichen Experten für das allgemeine englische Wortschatzwissen zu Websters Zeit stellen können? Der Einsatz von hoch angesehenen britischen Professoren als Experten hätte mit Sicherheit vernünf-

tig erscheinen können und zukünftige Ansätze erahnen lassen, die auch heutzutage noch in der Arbeits- und Organisationspsychologie eingesetzt werden, um sog. Situational Judgment Tests zu entwickeln. Diese Meinungen hätten jedoch in eine akademische Richtung verzerrt sein können. Es mag in diesem Zusammenhang von Interesse sein, dass Webster, der auch ein überzeugter Patriot und Anhänger der amerikanischen Revolutionsbewegung war, Definitionen für speziell nordamerikanische Ausdrücke wie „hickory“ und „skunk“ in sein Wörterbuch aufnahm. Er vereinfachte auch die Rechtschreibung in einer Weise, die eher in Einklang mit Benjamin Franklin stand, indem er zum Beispiel „centre“ durch „center“ und „musick“ durch „music“ ersetzte.[1] Britische Professoren des ausgehenden 18. Jahrhunderts an den Universitäten von Oxford und Cambridge wären wohl nicht von solchen Innovationen begeistert gewesen (sie hätten sie vermutlich überhaupt nicht akzeptiert), und dieser erwartete Widerstand hätte uneindeutige Resultate hervorgebracht, wenn gerade sie als inhaltliche Experten eingesetzt worden wären. Es wäre also vernünftiger (und im Sinne Thomas Jeffersons!), eine repräsentative Stichprobe von englischsprachigen Kolonisten heranzuziehen und Richtlinien zur Identifizierung geeigneter Antworten für Wortschatzitems zu entwickeln. Kurzum, wenn wir heute einen Vokabeltest ohne die Hilfe eines Wörterbuchs erstellen müssten, erschiene es uns als ein vernünftiges Vorgehen, ein breites Spektrum gebildeter Erwachsener zu befragen, die uns als Experten dienen.

Diese Überlegungen veranschaulichen, dass Wissensbereiche existieren können, die größtenteils auf (unter Umständen dogmatischen) Ansichten einzelner fußen und keine anderen objektiven Standards zur Verifikation aufweisen als gesellschaftliche Meinungen und Interpretationen. Dennoch können diese Wissensbereiche wichtige Informationen bezüglich der Fähigkeiten einer Person liefern; schließlich lädt Wortschatzwissen in der Regel sehr hoch auf dem *g*-Faktor der allgemeinen Intelligenz (Carroll, 1993). Für solche Wissensbereiche könnte es eine conditio sine qua non sein, sozial geteilte Wissensstandards zur Evaluation individueller Antworten zu verwenden. Dieses Konzept stellt Wissen in einen Zusammenhang mit Erfahrung und Ansichten und ist in den Schriften von eminenten Philosophen wie Plato und John Stuart Mill verwurzelt. Darüberhinaus ist natürlich die Akzeptanz einer von den meisten Mitgliedern einer Gesellschaft geteilten Ansicht ein essentieller Bestandteil demokratischer Einrichtungen!

Die Bestimmung von Wissen in „weichen“ und im Entstehen begriffenen Bereichen wie emotionaler und sozialer Intelligenz, in denen die Kodifizierung und die Formalisierung von Wissen gerade erst beginnt, verlangt geradezu nach der Anwendung dieser neuen Technologien. Diese unscharf definierten Bereiche sind häufig mit weitreichenden Konsequenzen verbunden: Beispielweise haben Wissen und Expertise bezüglich Fahrsicherheit, Führung oder sozialer Beziehungen erheblichen Einfluss auf die individuelle Lebensqualität. Für diese Diskussion ist es hilfreich, sich vor Augen zu führen, dass diese Wissensbereiche analog zu der Situation sind, die unser Wortschatztestentwickler des 18. Jahrhunderts erlebt hätte, da für sie ebenfalls keine objektiven und umfangreichen Informationen verfügbar sind und – dies ist genauso wichtig – die Identifizierung von geeigneten inhaltlichen Experten problematisch ist. Zur Erfassung der für diese Bereiche entwickelte Maße würden die Konsistenz der kognitiven Strukturen eines Individuums mit Hilfe eines Auswertungsstandards bewerten, der mit einem Gruppenkonsens korrespondiert. Sie wären daher in ähnlicher Weise zu interpretieren.

[1] `http://www.m-w.com/`

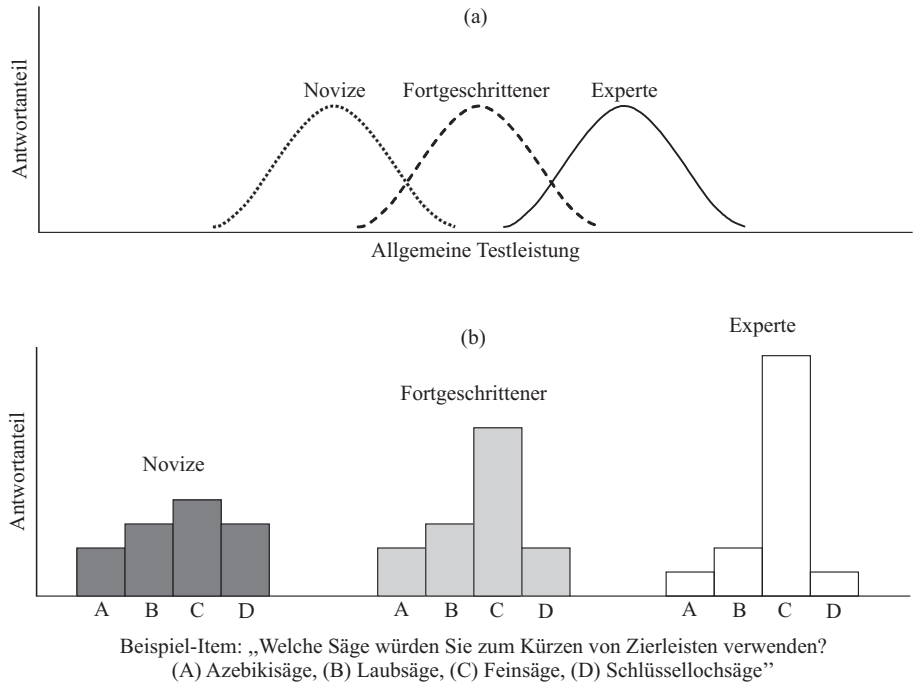

Abbildung 8.1 Leistung bei einem konventionellen Test auf Skalen- und Itemlevel über drei Expertisestufen. A: Allgemeine Leistungsverteilungen für die Skalen eines Multiple-Choice-Tests. B: Theoretische Antwortverteilungen für ein Multiple-Choice-Item, bei dem C die richtige Antwort ist.

8.3 Wissen, Antwortverteilungen und Expertise

Unsere Konzeptualisierungen von KBM entstanden aus Erwartungen bezüglich der Art und Weise, mit der sich Item-Antwortverteilungen als eine Funktion der Expertise in einer Stichprobe ändern könnten. Es wird gemeinhin angenommen, dass Wissen mit größerer Expertise in einem bestimmten Bereich wächst. Wenn eine Stichprobe von Auszubildenden über die Zeit beobachtet und wiederholt mit Standard-Wissensitems als Novizen, Gesellen und Fachleute untersucht würde, könnten daher die Antwortverteilungen, die in Abbildung 8.1a dargestellt sind, ihren Zuwachs an Expertise beschreiben. Die Verteilungen in Abbildung 8.1a veranschaulichen die interindividuellen Unterschiede ebenso wie den erwarteten Zuwachs an Wissen. Wie Abbildung 8.1b zeigt, würden mit zunehmender Expertise größere Anteile an Probanden die „richtige" Antwort für ein einzelnes Testitem wählen.

Nehmen wir jedoch einmal an, eine Stichprobe von Studenten, die sich mit EI befassen, würde mit szenariobasierten Items untersucht, die eine Bewertung auf einer Likert-Skala erfordern. Die Probanden könnten beispielsweise gebeten werden, ihre Zustimmung zu der folgenden Aussage einzuschätzen: „EI kann als Wissensschatz eines Individuums über die soziale Welt definiert werden". Ähnliche Aussagen wurden zur Definition sozialer Intelligenz vorgeschlagen (siehe Cantor & Kihlstrom, 1987), aber nicht für emotionale Intelligenz. Für diese Art der Fragestellung könnten sich die mit größerer Expertise assoziierten Item-Antwortverteilungen sowohl in der zentralen Tendenz als auch in der Varianz unterscheiden. Eine Veränderung in der zentralen Tendenz könnte auftreten, wenn Studenten lernen, dass einige EI-Konzeptualisierungen Implika-

tionen für soziales Wissen beinhalten. Veränderungen in der zentralen Tendenz dieser Antwortverteilungen werden in Abbildung 8.2a dargestellt.

Eine Varianzverringerung könnte auch auftreten, wenn Studenten ein genaueres Verständnis von emotionaler Intelligenz entwickeln und erkennen, dass EI-Konzeptualisierungen zwar auch Implikationen für soziales Wissen beinhalten, sich aber auf Emotionskonstrukte konzentrieren (siehe Kapitel 2 von Neubauer & Freudenthaler). Abbildung 8.2b zeigt eine Varianzverringerung in den Antwortverteilungen, die mit erhöhter Urteilsgenauigkeit einhergeht.

Diese beiden Trends sind für das Verständnis des Wachstums und der Verbesserung von Wissen durch Reflexion, Erfahrung und formale Bildung von allgemeiner Bedeutung. Per Definition fehlt naiven Personen jegliche Grundlage zum Verständnis von Beziehungen oder Ereignissen, und ihre Reaktionen, die oft wenig ausgereifte Überlegungen widerspiegeln, sind vielleicht unvernünftig. Manchmal zeigen sie sich sogar ignorant gegenüber grundlegenden Beziehungen zwischen Ereignissen oder übertreiben deren Bedeutung. Mit zunehmender Entwicklung werden sich Personen ihres Verständnisses von Beziehungen und Ereignissen jedoch bewusster, und sie werden genauer in ihren Einschätzungen. Es lohnt sich zu bedenken, dass, in dem Ausmaß, in dem schlechte Leistungen in einem Wissenstest als Fehler angesehen werden können, die Antworten von Nicht-Experten variabler sind als die von Experten und möglicherweise auch eine andere zentrale Tendenz aufweisen.

Diese Konzeptualisierung zeigt, dass untersuchte Personen mit ansteigender Leistung (Expertise) und somit geringerer Fehlerrate sowohl bei konventionellen als auch bei szenariobasierten Testitems tendenziell besser übereinstimmen. Die zentrale Tendenz von Experten-Antwortverteilungen für einzelne szenariobasierte Items sollte in etwa gleich sein wie die zentrale Tendenz der Antwortverteilungen von Nicht-Experten (z. B. Gesellen), wenn das Wachstum des Wissens über die Expertisestufen in erster Linie mit Veränderungen in der Varianz verbunden ist (Abbildung 8.2b). Diese Beobachtung ist ebenfalls auf konventionelle Multiple-Choice-Items anwendbar (Abbildung 8.1b), aber sie ist hier von geringem praktischen Wert, weil die Formulierung von guten Multiple-Choice-Items die a priori-Kenntnis der richtigen Antwort erfordert. Szenariobasierte Items erfordern nicht immer, dass die richtige Antwort spezifiziert werden kann oder überhaupt bekannt ist.

Dennoch ist es ebenso möglich, dass erhöhte Expertise mit Veränderungen der zentralen Tendenz und Varianz einhergeht. Dieses „mittlere“ Modell wird in Abbildung 8.2c dargestellt. Bis jetzt vermögen wir nicht zu sagen, welche Arten von Items mit ansteigender Expertise Veränderungen über die Expertisestufen in Varianz, zentraler Tendenz oder auch beidem zeigen. Aber wir machen trotzdem auf diese Möglichkeit aufmerksam und zeigen an, dass zukünftige Arbeiten zu KBM diese Beziehungen untersuchen sollten.

Eine sehr weitreichende Implikation von erfolgreich angewendeten konsensbasierten Tests und Inventaren ist die Rechtfertigung und Bestätigung von allgemeinen demokratischen Prozessen, welche die „Tyrannei der autokratischen Expertise“ stürzen. Die Untersuchung der hypothetischen Antwortverteilungen vieler Novizen gegen die einer Handvoll Experten sollte flachere Verteilungen aufdecken, die sich leichter an veränderliches Weltwissen anpassen und sich mit diesem verändern. Nimmt man an, dass die Korrelation zwischen dem Wissen von Novizen und Experten in diesen Instrumenten durch den Schnittpunkt ihrer Korrelationen mit einer allgemeineren „Wahrheit“ ver-

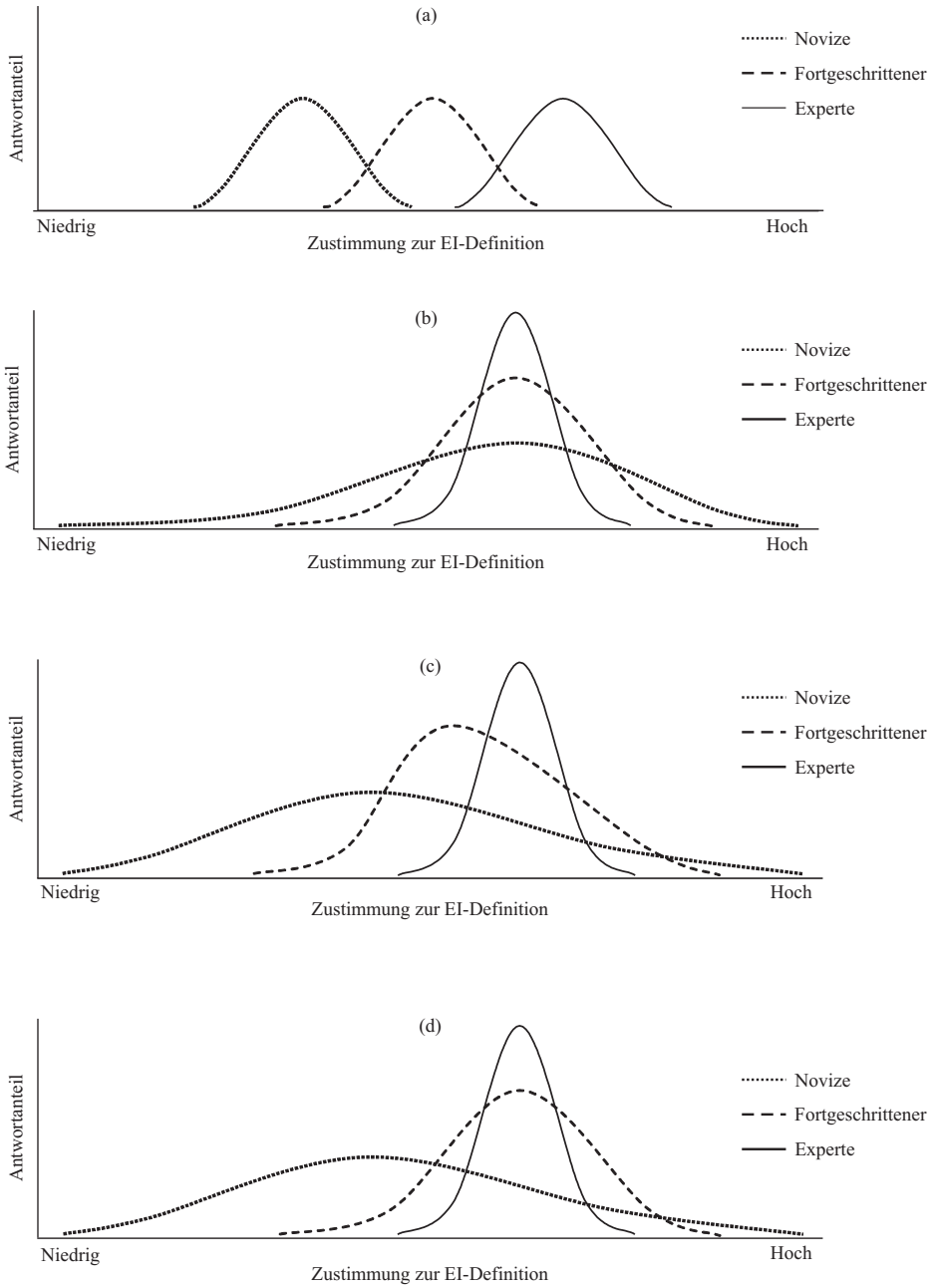

Abbildung 8.2 Item-Antwortverteilungen über drei Expertisestufen, welche mit szenariobasierten Items korrespondieren, die eine Antwort auf einer Likert-Skala erfordern (z. B. *EI kann definiert werden als individuelles Wissen über die soziale Umwelt*). A: Antwortverteilungen mit veränderter zentraler Tendenz und gleicher Varianz. B: Antwortverteilungen mit gleicher zentraler Tendenz und unterschiedlicher Varianz. C: Erwartete Antwortverteilungen mit unterschiedlichen zentralen Tendenzen und unterschiedlichen Varianzen. D: Beobachtete Itemverteilungen für szenariobasierte Items.

mittelt wird,[2] mag es sich herausstellen, dass verschiedene Gruppen von Novizen die Wahrheit genauer reflektieren als Experten. Zumindest ist dies eine Hypothese, die ebenfalls untersucht werden sollte. Einige Implikationen dieser Beziehungen werden weiter unten dargestellt.

8.4 Situational Judgment Tests zur Messung emotionaler Intelligenz

Wir haben oben bereits festgestellt, dass SJTs ideal zur Beobachtung von Veränderungen in Item-Antwortverteilungen über verschiedene Expertisestufen hinweg geeignet sind, sowohl in Hinsicht auf die zentrale Tendenz als auch die Varianz. Wir beschreiben dabei SJTs allgemein als Tests, die:

1. entweder implizit oder explizit ein Szenario beschreiben, das ein Ereignis, eine Situation oder einen Prozess simuliert oder abbildet. Die Szenarien können die Reaktion auf Probleme (die Lösungen erfordern), die Beibehaltung von Erfolg oder die Interpretation von Ereignissen darstellen. Ein Verständnis dieser Beschreibungen mag die Anwendung von Wissen erfordern, das entweder durch Erfahrung oder formal erworben wurde.
2. eine Liste von Alternativen, die mit jedem Szenario in Verbindung stehen, bereitstellen. Diese Listen können handlungsorientierte oder interpretative Alternativen enthalten oder den untersuchten Personen erlauben, in einem offenen Antwortformat ihre Meinung und ihr Wissen zu beschreiben.
3. die untersuchten Personen dazu verpflichten, entweder die Alternativen, die mit den Szenarien in Verbindung stehen, zu bewerten (z. B. die Angemessenheit der Alternativen zu bewerten) oder neue Alternativen und Analysen der Szenarien im Falle eines offenen Antwortformats zu generieren.

Die Leistung in SJTs wird durch eine Analyse der Probandenantworten quantifiziert. Ein SJT kann viele Szenarien, die jeweils wie Items behandelt werden, beinhalten oder ein einzelnes Szenario beschreiben, bei dem die einzelnen Alternativen wie Items behandelt werden. Wie die Probandenurteile bezüglich eines Standards ausgewertet und wie diese Standards entwickelt werden, ist das Thema der folgenden Abschnitte.

Die Bewertungsstandards der meisten existierenden situativen SJTs beruhen darauf, dass inhaltliche Experten die Alternativen zu jedem Szenario beurteilen bzw. einschätzen (vgl. McDaniel et al., 2001). Diese Daten werden dann benutzt, um die Expertenauswertungsschlüssel zu konstruieren, zum Beispiel durch Berechnung der gemittelten Expertenbewertungen für jede Alternative. Anschließend wird der Prozentsatz richtiger Antworten, ein Abweichungsmaß oder eine Korrelation der Probanden-Ratings mit dem Auswertungsstandard errechnet, um die Probandenantworten im Vergleich zu den expertenbasierten Standards zu evaluieren. In Übereinstimmung mit der Verwendung einer Reihe von Verfahren zur Bewertung der Leistung in SJTs mit expertenbasierten Auswertungsstandards können verschiedene Verfahren eingesetzt werden, um SJTs konsensbasiert zu scoren. Tabelle 8.1 stellt diese verschiedenen Methoden dar.

Während in diesen Ansätzen implizit die klassische Testtheorie zum Einsatz gelangt, ist es vernünftig, ebenfalls IRT-Analysen bei Vorliegen ausreichender Datenmengen durchzuführen. Ein zur Bewertung impliziten Wissens im Straßenverkehr entwickelter SJT wird in Tabelle 8.2 zur Veranschaulichung dieses Ansatzes dargestellt.

[2]Das heißt: $R(K_n, K_e) = R(K_n, K_t) \bigcap R(K_e, K_t)$.

Tabelle 8.1 Ansätze zur Bewertung von SJTs

Methode	Anwendung	Autoren
Prozentuale Übereinstimmung: Likert-skalierte Daten werden erhoben; die einzelnen Antworthäufigkeiten werden für die Gewichtung der Optionen genutzt.	Emotionale Intelligenz	Mayer et al. (2003)
Einfache Distanz: Mit Likert-skalierten Daten werden Itemmittelwerte über alle Probanden berechnet. Die absoluten Abweichungen der individuellen Ratings von den einzelnen Itemmittelwerten sind die einfachen Distanzen. Die Probandenleistung wird als mittlere Itemdistanz quantifiziert.	Wissen im Straßenverkehr	Legree et al. (2003)
Standardisierte Distanz: Ähnlich wie die einfache Distanz-Methode, jedoch werden die Ratings zunächst transformiert, um sie innerhalb der Antworten eines Probanden zu standardisieren. Dieser Ansatz kontrolliert Antworttendenzen einzelner Probanden in bestimmten Subsegmenten der Skala.	Soziale Intelligenz & *g*	Legree (1995); Legree et al. (2000)
Quadrierte Differenz: Ähnlich wie die einfache Distanz-Methode, aber die Itemwerte sind die quadrierten Differenzen. Verleiht größeren Differenzen zusätzliches Gewicht.	Implizites Wissen	Sternberg et al. (2000)
Korrelation: Die Korrelation zwischen individuellen Beurteilungen und gemittelten Urteilen quantifiziert die Leistung.	Führung	Psotka, Streeter, Landauer, Lochbaum & Robinson (2004)

Diesem Wissenstest lag ein Modell der Fahrleistung zugrunde, welches annahm, dass Fahrer durch Anpassung ihrer Geschwindigkeit an gefährliche Verkehrsbedingungen ihr Verkehrsrisiko verringern können (Legree et al., 2003). Dieser SJT wurde mittels der Berechnung von Distanzwerten zwischen den Probandenantworten und dem Auswertungsstandard für jedes der insgesamt 14 Items ausgewertet.

Die meisten SJTs wurden für die Anwendung in Organisationen entwickelt. Diese Skalen offerieren gewöhnlich arbeitsbezogene Problemszenarien und die Probanden werden instruiert, zwischen möglichen *Lösungsalternativen* zu wählen. Im Gegensatz dazu präsentieren die Skalen der MEIS und des MSCEIT – und wohl auch die sog. Conditional Reasoning Tests (siehe James, 1998) – bestimmte Informationen und fordern von den Probanden, zwischen möglichen *Interpretationen* zu wählen. Folglich können die Skalen in MEIS und MSCEIT als abstrakter angesehen werden als Standard-SJT-Maße, weil sie vermittelnde kognitive Prozesse untersuchen, die EI zugrunde liegen, im Gegensatz zur einfachen Simulation beobachtbarer Entscheidungen.

Für den vorliegenden Beitrag ist relevant, dass SJTs im Prinzip auch auf andere nichttraditionelle Arten entwickelt werden könnten, um zusätzliche EI-Aspekte (oder ein beliebiges anderes Konstrukt) zu erläutern. Es könnte zum Beispiel ein SJT konstruiert werden, in dem bestimmte Informationen dargestellt werden. Die Zeit, die Probanden brauchen, um einfache Aussagen oder nonverbale Stimuli zu beurteilen, würde

Tabelle 8.2 Wissenstest zu sicheren Fahrgeschwindigkeiten

Nehmen Sie an, jemand fährt ein sicheres Auto in leichtem Verkehr unter optimalen Bedingungen. Bitte *schätzen* Sie das Ausmaß ein, in dem diese Person (der Fahrer) unter den im Folgenden genannten Bedingungen die Geschwindigkeit verändern und langsamer werden sollte (oder dies nicht tun sollte), um seine Sicherheit zu garantieren.

Bedingungen:	−20 MPH Langsamer werden	−10 MPH	0 MPH Gleiche Geschw.
1. Schnee und starker Verkehr			
2. Klare Sicht und leichter Verkehr			
3. Schnee und kein Verkehr			
4. Trockene Straßen um Mitternacht			
5. Stress wg. Arbeitsproblemen			
6. Mittelstarker Verkehr			
7. Schotter & leichter Verkehr			
8. Freie Straßen und etwas windig			
9. Leichter Regen und kurvenreiche Straßen			
10. Verärgert und leichter Regen			
11. Leichter Verkehr und hügeliges Gelände			
12. Leicht abgefahrene Reifen			
13. Familienärger wg. Finanzen/Geld			
14. Krank mit einer Kopfgrippe			
	−20 MPH Langsamer werden	−10 MPH	0 MPH Gleiche Geschw.

auf auf ähnliche Weise wie bei Reaktions- oder Kontrollzeitaufgaben geschätzt werden (siehe Detterman, Caruso, Mayer, Legree & Conners, 1992). Ein solcher SJT würde die Verarbeitungsgeschwindigkeit messen, die mit EI-Kognitionen verbunden ist und mit entsprechenden Chronometrie- und Speedkonzeptionen übereinstimmen (siehe Carroll, 1993; Jensen, 1998).

8.5 Konsensbasierte Messung: Empirische Befunde

8.5.1 Ein SJT zur Messung der sozialen Intelligenz für Supervisoren

In früheren Arbeiten mit SJTs (Legree, 1995; Legree & Grafton, 1995) evaluierten wir unsere Konzeptualisierung der Wissensentwicklung durch den Vergleich von expertenbasierten Auswertungsstandards (die die mittleren Ratings einer kleinen Anzahl von Experten darstellten) mit den Antworten von „echten“ Probanden (ebenfalls gemittelte Ratings). In unserem SJT für Supervisoren wurden 49 Szenarien beschrieben und eine Gesamtzahl von 198 Alternativen mit jeweils drei bis fünf Alternativen pro Szenario aufgelistet. Jedes Szenario beschrieb ein zwischenmenschliches Problem und bot mögliche Lösungsalternativen dafür an. Die untersuchten Personen und die Experten schätzten für jedes Szenario die Angemessenheit aller in den Alternativen beschriebenen Handlungen ein. Wir berechneten dann die über die Probanden gemittelten Ratings für jede der 198 Alternativen. Dabei beobachteten wir eine hohe Korrelation zwischen dem expertenbasierten Auswertungsstandard und dem mittleren Item-Rating der Probanden von $r = .72$, ($N = 198, p < .001$) und eine sehr hohe, doppelt minderungskorrigierte (d. h., Korrektur für Unreliabilität sowohl für die Experten- als auch für die Probandenurteile) Korrelation von $r = .95$.

Zu Beginn hatten wir erwartet, dass die mittleren Probandenratings nur eine grobe Schätzung des Expertenstandards liefern würden. Wir hatten gehofft, dass diese Schätzung durch eine moderate Korrelation zwischen den Mittelwerten belegt würde, die im Bereich von .40 bis .60 liegen würde, und hatten geplant, eine rekursive Prozedur zu verwenden, um der Reihe nach Gruppen von Individuen mit jeweils höheren Wissenslevels zu identifizieren. Diesen Ansatz wollten wir nutzen, um Tests auszuwerten, für die Expertenmeinungen nicht verfügbar waren. Informationen aus ausgewählten Probandengruppen wäre dann dazu verwendet worden, um validere Auswertungsstandards für den Supervisoren-SJT zu entwickeln, die sich stärker an die Expertenstandards annähern würden. Diese Standards würden dann als *konsensbasierte Standards* bezeichnet und der Prozess als KBM.

Auf Grundlage der beobachteten und korrigierten Korrelationen zwischen den mittleren Probanden- und Expertenratings, .72 und .95, wurde die Verwendung rekursiver Prozeduren zur Verfeinerung des Scoringmusters als nicht notwendig betrachtet. Wir berechneten zusätzlich auch Probandenscores mittels Verwendung zweier verschiedener Standards, die auf den Experten- und Probandenmittelwerten basierten, und korrelierten dann zwei Arten von Scores; die Korrelation betrug $r = .88$ ($N = 198; p < .001$).

Diese Korrelationen bestätigten, dass die gemittelten Ratings der Probanden einen alternativen Auswertungsstandard für den SJT lieferten, und diese Erkenntnis warf Fragen bezüglich der Angemessenheit der beiden Standards auf. Wir kamen zu dem Schluss, dass der probandenbasierte Standard vorzuziehen war, weil diese Werte aufgrund der großen Personenzahl ($N = 193$) reliabler waren als der Expertenstandard. Wir wandten diese Methode dann an, um zwei zusätzliche Skalen zur sozialen Intelligenz auszuwerten, für die Expertenmeinungen nicht verfügbar waren. Eine konfirmatorische Faktorenanalyse der Werte dieser drei Skalen und herkömmlichen Fähigkeitstestbatterie (die Armed Service Vocational Aptitude Battery) wies die Existenz eines separaten, zu unserem Modell der sozialen Intelligenz korrespondierenden g-Faktors nach (Legree, 1995).

8.5.2 Anwendungen zur Erfassung von *g* und Verkehrssicherheitswissen

Obwohl unser Modell der sozialen Intelligenz bestätigt werden konnte, waren wir der Ansicht, dass sich KBM auch in anderen Bereichen bewähren sollte. Dies sollte durch die Validierung konsensbasierter Werte anhand konzeptuell relevanter und wichtiger Kriterien und durch die Übereinstimmung zwischen auf Probanden- und Expertenmeinungen beruhenden Testwerten geschehen. Daher untersuchten wir in späteren Studien die Vielseitigkeit dieser Methode, indem wir zwei Arten von Skalen entwickelten und validierten: sechs Wissenstests zur Messung allgemeiner kognitiver Fähigkeiten (g) auf der einen und zwei Tests zur Erfassung impliziten Wissens über sicheres Fahrverhalten im Straßenverkehr auf der anderen Seite. Die meisten dieser Instrumente (es gab eine Ausnahme) erforderten von den Personen Antworten auf Likert-skalierte Items, zum Beispiel die Einschätzung der Häufigkeit von in mündlicher Konversation gebrauchten Wörtern und Ausdrücken oder die Einschätzung des Ausmaßes, in dem Fahrer ihre Geschwindigkeit in Anbetracht von Unfallrisiken mäßigen sollten. Diese Skalen dienten der Abbildung inzidentellen Lernens und impliziten Wissens zur Vorhersage und zum Verständnis menschlicher Leistung. Solches Wissen und damit assoziierte Expertise wird normalerweise nur langsam und inkrementell als Ergebnis von gesammelter und reflektierter Erfahrung erworben (Sternberg et al., 2000). Zur Entwicklung von Auswertungsstandards für diese Tests konnten weder eine objektive Wissensgrundlage noch Experten herangezogen werden. Somit konnte die Leistung auf diesen Skalen nur mit Hilfe konsensbasierter Scoringalgorithmen evaluiert werden.

Die g-Testbatterie wurde einer ausgelesenen, aus Air-Force-Rekruten bestehenden Militärstichprobe vorgegeben. Die Faktorwerte, die aus dieser experimentellen Testbatterie extrahiert wurden, korrelierten zu .54 mit den aus einer konventionellen Testbatterie extrahierten Faktorwerten (d. h. dem g-Faktor der Intelligenz). Die aufgrund der starken Varianzeinschränkung in der vorliegenden Stichprobe minderungskorrigierte Korrelation betrug .80 (Legree et al., 2000). Fünf der sechs experimentellen Skalen korrelierten ebenfalls signifikant mit g. Eine Parameterschätzung von .80 ist typisch für Korrelationen, die zwischen verschiedenen IQ-Testbatterien erzielt werden (vgl. Carroll, 1993). Eine konfirmatorische Faktorenanalyse der korrigierten Korrelationsmatrix lieferte einen Pfadkoeffizienten von .97 zwischen den beiden latenten Faktoren (unsere g-Tests und die konventionelle Testbatterie). Somit waren wir also in der Lage, eine sehr hoch auf dem g-Faktor ladende Testbatterie zu erzeugen, deren Scoring ohne inhaltliche Experten oder objektives Wissen, sondern stattdessen nur durch Konsens von Nicht-Experten erfolgen konnte.

Die Tests zur Erfassung des impliziten Straßenverkehrswissens wurden an Soldaten erprobt, von denen auch Daten bezüglich der Beteiligung an Autounfällen erhoben wurden. Verglichen mit den üblicherweise eingesetzten Leistungsbereichen erscheint die Verwicklung in Autounfälle ein eher ungewöhnliches Maß zu sein, das zudem in Metaanalysen nur äußerst geringe Zusammenhänge mit Wissens-, Fertigkeits- und Fähigkeitsmaßen, einschließlich der Intelligenz, aufweist (Arthur, Barrett & Alexander, 1991; Veling, 1982). Wie Tabelle 8.3 jedoch zu entnehmen ist, korrelieren beide Tests zur Erfassung des impliziten Straßenverkehrswissens zwischen $-.11$ und $-.20$ signifikant mit Kriterien der Verwicklung in Unfälle (Legree et al., 2003).

Obwohl diese Werte gering erscheinen mögen, übertreffen sie Koeffizienten, die sich gewöhnlich für stabile Merkmale einstellen, und sie können dazu genutzt werden, die Sicherheit im Straßenverkehr zu erhöhen. Die von uns erzielten Werte demonstrieren daher die Nützlichkeit der Verwendung konsensbasierten Scorings bei der Erfassung impliziten Wissens auch in diesem exotischen Anwendungsbereich.

Tabelle 8.3 Itemantwortverteilungen bezüglich des Wissens zu sicheren Fahrgeschwindigkeiten und Faktorladungen

Test Items	*M*	*SD*	% Speed[a]	Faktorladungen: Emotionales Wissen	Unkompliziert	Niederschlag
Ärger über Familienfinanzen	8.39	5.30	16	**.73**	−.01	.04
Krank mit Kopfgrippe	8.50	5.08	13	**.73**	−.09	−.05
Leicht abgefahrene Reifen	7.62	4.92	3	**.55**	.05	−.02
Von der Arbeit gestresst	7.61	4.90	18	**.46**	.06	.17
Leichter Verkehr & hügelig	6.17	4.28	20	**.44**	.17	.05
Klare Sicht & leichter Verkehr	1.50	3.19	78	−.02	**.92**	−.11
Freie Straßen & windig	2.77	3.52	49	−.02	**.68**	.14
Trocken & Mitternacht	4.52	3.79	28	.20	**.62**	.02
Leichter Regen & Kurven	10.40	4.07	1	.06	−.08	**.59**
Verärgert & leichter Regen	10.59	4.43	3	.23	−.12	**.44**
Schnee & kein Verkehr	11.17	4.11	2	.02	.01	**.49**
Schnee & starker Verkehr	14.81	4.01	0	−.07	.13	**.40**
Mittelstarker Verkehr[b]	7.70	4.17	8			
Schotter und leichter Verkehr[b]	7.77	4.02	7			

	Kriterienkorr.			Faktorkorr.		
	Fehler-Rate	Fehler-Status[c]	*g*			
Emotional	−.19‡	−.20‡	.10*	1.00	.19	.50
Trockenes Wetter	−.10*	−.16†	.31‡		1.00	.25
Niederschlag	−.16‡	−.16†	.11†			1.00

Anmerkungen. $N = 387$. [a]Prozent an Probanden, die angaben, dass die Geschwindigkeit nicht reduziert werden solle. [b]Wegen Nebenladungen aus der Analyse ausgeschlossene Variablen. [c]Fehlerstatus bei $N = 211$.
* $p < .10$. †$p < .05$. ‡$p < .01$.

Der Straßenverkehrswissenstest zu sicheren Fahrgeschwindigkeiten wird in Tabelle 8.2 dargestellt und ist eine von zwei Skalen, die von uns zur Erfassung impliziten Wissens in diesem Bereich entwickelt wurden. Es bestätigt die Bedeutung der mit emotionaler Intelligenz verbundenen Konstrukte, dass bei der Faktorisierung der Items einer der drei Faktoren durch emotional und internal bedeutsame Items definiert werden konnte (siehe Tabelle 8.3). Obwohl dieser Faktor nur minimal auf dem *g*-Faktor lud, war er für die Kriterien zur Unfallverschuldung in hohem Maße prädiktiv. Diese Daten zeigen, dass sich sichere Verkehrsteilnehmer in höherem Maße der Wichtigkeit angemessener Geschwindigkeit bewusst sind, wenn sie emotionalem oder internalem Stress ausgesetzt sind.

Natürlich hätten wir auch verschiedene Personen als „Experten“ bezeichnen können, um die Auswertungsstandards für die durch die *g*-Testbatterie und die beiden Tests zur Erfassung des impliziten Straßenverkehrswissens abgedeckten Wissensbereiche zu entwickeln. Jedoch glauben wir, dass alle Expertenaussagen, die für Items aus diesen Bereichen zu erwarten sind, verdächtig sind. Willkürlich ernannte Experten, die nicht mehr echte Expertise aufwiesen als die Probanden in unserer Studie, hätten sich unserer Meinung nach in ihren Antworten nicht *quantitativ* von unseren Probanden unterschieden, sondern *qualitativ*: Da sie bei ihren Antworten einen besonderen Zweck verfolgt

hätten, wären diese qualitativ verschieden gewesen (so hätten sie nicht auf der Grundlage impliziter Kenntnisse geantwortet, sondern explizite, idiosynkratische Informationen verwendet). Wissen über die Verwendungshäufigkeit bestimmter Wörter in mündlichen Konversationen und über Straßenverkehrssicherheit sind Beispiele für Bereiche, in denen Erfahrung eine Voraussetzung für bessere Leistung ist, denen es aber an echten Experten mangelt.

Der Implicit Association Task-Test, der den einzigen experimentellen Test darstellt, für den keine Likert-Antwortskala verwendet wird, ist sogar noch spezieller. Er erfasst die Fähigkeit eines Probanden, binäre Muster zu verstehen (siehe Psotka, 1977), und jedes Item erfordert die Weiterführung von Mustern aus Folgen von X und O (z. B. XOXOXO?). Hierzu konnte kein anderer Auswertungsstandard gefunden werden, weil die als Stimuli verwendeten Muster nicht in Übereinstimmung mit vorher spezifizierten Regeln oder Beziehungen definiert wurden, die die korrekte Antwort bestimmt hätten. Folglich konnten diese Items nur konsensbasiert ausgewertet werden. Nichtsdestotrotz korrelierte die Leistung bei dieser Aufgabe mit dem *g*-Faktor.

8.5.3 Zusätzliche empirische Unterstützung der Befunde zu Experten- und Probandenvergleichen

Die obigen Daten demonstrieren die Effektivität des Ansatzes der KBM, insbesondere was ihre prädiktive Validität und Möglichkeiten zur Generierung nützlicher Auswertungsstandards anbelangt. Es besteht wenig Zweifel darüber, dass KBM zum Antwortenscoring bei psychologischen Tests verwendet werden kann, insbesondere für Tests, die für exotische, „weiche“ Wissensbereiche, in denen es kaum objektives Wissen gibt und die entweder sehr hoch oder sehr niedrig auf dem *g*-Faktor laden, entwickelt wurden. Unsere Konzeptualisierung sagte auch eine hohe Korrelation zwischen experten- und konsensbasierten Auswertungsstandards voraus, und natürlich auch für die Testwerte, die auf Basis dieser Standards gewonnen werden. In der Untersuchung zur anfänglichen Bewertung unseres Modells korrelierten zum Beispiel für den Supervisoren-SJT die Experten- und die konsensbasierten Auswertungsstandards zu .72 und die auf diesen Standards basierenden Testwerte zu .88. Weil Experten oft schwer zu finden und wenn gefunden, dann teuer sind, hat ein großer Teil der Forschungsergebnisse, die auf expertenbasierten Maßen basieren, eine niedrige Reliabilität. Uns sind drei andere Studien bekannt, die aus einer großen Expertengruppe gewonnene expertenbasierte Standards einsetzten und die benötigten Reliabilitätsanforderungen somit aller Voraussicht nach erfüllen. Es gibt mit Sicherheit weitere Studien, die diese Analysen unterstützen könnten, aber die Daten von Testprobanden werden selten zur Schätzung von Expertenbeurteilungen benutzt.

Der 'nicht beförderte Offizier'-SJT. Dieser SJT wurde zur Evaluation der supervisorischen Fähigkeiten ranghöherer Soldaten entwickelt. Er beschreibt 71 Problemszenarios und führt 362 Handlungsmöglichkeiten auf. Um die konsensbasierte Scoringmethode zu evaluieren, wurden die Antwortprotokolle sowohl auf Basis experten- ($N = 88$) als auch konsensbasierter Standards ($N = 1891$) ausgewertet (Heffner & Porr, 2000, W. B. Porr, persönliche Kommunikation, Juli 2003). Die Gesamtleistungswerte korrelierten zu .95 und die Auswertungsstandards zu .89.

Studien zur emotionalen Intelligenz. Der MSCEIT (Mayer et al., 2003; siehe auch Kapitel 2 und 7 des vorliegenden Buchs), der wohl die am weitesten entwickelte leis-

tungsbasierte Testbatterie für emotionale Intelligenz ist, stellt sowohl experten- als auch konsensbasierte Normwerte zur Verfügung. Die Expertengruppe bestand aus 21 Mitgliedern der „International Society for Research on Emotions“ und das konsensbasierte Scoring beruhte auf den Antworten von 2112 Probanden. Die Korrelation zwischen den beiden auf diese Weisen erhaltenen Testwerten betrug .98 und die Auswertungsstandards korrelierten zu .91 miteinander. Die Forscher berichteten ebenfalls Interraterreliabilitäten (κ) für die Experten und zwei Teilstichproben von Nicht-Experten: Die Expertenreliabilitäten waren – wie von einem Modell mit verringerter Varianz bei zunehmender Expertise angenommen – konsistent höher als die Probandenreliabilitäten ($\kappa = .43$ versus $\kappa = .31/\kappa = .38$, $p < .01/p < .05$).

Studien zu implizitem Wissen über militärische Führung. Die dritte Art von Studien rekrutiert ihre Daten aus der Skala „Tacit Knowledge for Military Leadership“(TKML Hedlund et al., 2003; Psotka et al., 2004). Die TKML wurde zur Messung des praktischen und handlungsorientierten Wissens entworfen, das Führungspersonal in der US Army typischerweise durch Berufserfahrung gewinnt. Ausgangspunkt für die Entwicklung der TKML war, dass eine geordnete Expertise-Hierarchie bezüglich militärischer Führung erzeugt werden kann, indem die Werte von Obersten als Standard verwendet und mit denen von US Militärakademiekadetten (West Point), US Army Leutnants, Hauptmännern und Majoren verglichen werden. Die Skala wurde Soldaten vorgelegt (355 Kadetten, 125 Leutnanten, 117 Hauptmännern, 98 Majoren und 50 Obersten); die Obersten stellten dabei die Expertengruppe. Diese Gruppe beinhaltete die ranghöchsten Soldaten und solche, die am längsten im Militärdienst waren (mit einer durchschnittlichen Dienstzeit von 18 Jahren). Vergleiche zwischen den konsensbasierten Kadetten- und Experten-Auswertungsstandards (355 Kadetten und 50 Oberste) und ihren Testwerten lieferten sehr konsistente Ergebnisse. Die beiden Auswertungsstandards korrelierten zu .96 miteinander und die mit Hilfe dieser Standards errechneten Testwerte der Kadetten gar zu 1 (d. h. über .995). Ähnliche Ergebnisse wurden bei den Datenanalysen der Werte für die im Rang dazwischenliegenden Gruppen (Leutnants, Hauptmänner, Majore) gefunden.

Obwohl hohe Korrelationen zwischen experten- und konsensbasierten Standards bei der Validierung des Ansatzes helfen, hatten wir nahezu perfekte Werte doch kaum erwartet. Zudem war eine Verfeinerung der konsensbasierten Standards durch die Verwendung rekursiver Prozeduren für die von uns entwickelten Skalen wiederum nicht notwendig. Insgesamt gesehen legen diese Befunde eine Modifikation unserer Konzeptualisierung von KBM nahe: Der Hauptunterschied zwischen Fortgeschrittenen und Experten wird nunmehr durch eine zunehmende Genauigkeit der Schätzungen repräsentiert. Aus der Perspektive der Itemantwortverteilungen bedeutet dies geringere Varianzen um den Itemmittelwert. Der Übergang von Anfängern zu Fortgeschrittenen würde immer noch mit Veränderungen in den Antwortverteilungen und Mittelwerten einhergehen, da Anfänger keine oder nur eine geringe Erfahrungs- und Wissensbasis für ihre Antworten haben und ihre Antworten daher noch eher zufällig ausfallen. Dieses revidierte Modell ist in Abbildung 8.2d dargestellt. Um dieses Modell zu bewerten ist es notwendig, die Antwortverteilungen umfangreicher Stichproben von Personengruppen zu überprüfen, die in ihrem Grad an Expertise variieren.

Die meisten Datengrundlagen sind für diesen Zweck nicht geeignet, weil sich in den meisten nicht-geschichteten Stichproben sehr wenige Anfänger oder Experten befinden und die Identifikation solcher Personen sehr schwierig ist. Die TKML-Datenbasis ist hingegen einzigartig, weil sie eine große Anzahl Anfänger (355 Kadetten), Experten (50

Oberste) und Probanden fortgeschrittener Stufen (125 Leutnants, 117 Hauptmänner und 98 Majore) enthält. Diese Gruppen unterscheiden sich in einer Reihe salienter Dimensionen, die einen Einfluss auf Expertise haben, nämlich vor allem Alter, Erfahrung und Bildung. In der Tat haben Kadetten wenig militärische Berufserfahrung, aber sie haben – mit hoher Wahrscheinlichkeit – Erfahrungen in zwischenmenschlichen Beziehungen und Problemen, Fragen der Autorität, der Fürsorge und des Gehorsams; die Szenarien im TKML bedienen sich vor allem in diesen Bereichen. Obwohl sie also Anfänger sind, besitzen sie bereits relevantes Wissen.

Es sollte uns nun nicht mehr allzu sehr verwundern, dass bei der Korrelierung der Mittelwerte der Itemantwortverteilungen von 355 Kadetten mit den Mittelwerten von 50 Obersten (Experten) die Gesamtkorrelation ziemlich hoch ausfiel ($r = .96$) *und* die Steigung der Regressionsgeraden nahezu eins (.99) betrug. Die Steigung der Regressionsgeraden deutet auf ein vergleichbares Ausmaß an Varianz in den beiden Teilmengen von Itemmittelwerten hin. Selbst für einen so großen Expertise-Unterschied, wie er zwischen Kadetten (mit 0 Jahren Berufserfahrung) und Obersten (mit durchschnittlich 18 Jahren Erfahrung) vorherrscht, ist die Verwendung des Gruppendurchschnitts als Standard nicht von dem der Experten zu unterscheiden. Und trotzdem unterscheidet dieser Standard sauber zwischen den zwei Gruppen. Wie ist dies möglich? – Obwohl der Gesamtmittelwert für jede der Alternativen der Szenarien für Kadetten und Oberste praktisch derselbe war, schnitten auf der gesamten TKML-Skala sogar die besten 25% der Kadetten signifikant schlechter als die Obersten ab. Insgesamt gesehen lag der Mittelwert der besten 25% der Kadetten bei .73, wohingegen der Mittelwert der Obersten .82 betrug ($t = 4.27, 132\ df, p. < .01$). Dies entspricht einer Differenz von .36 Standardabweichungseinheiten und zeigt, dass konsensbasierte Standards tatsächlich das erfassen, was die Skala zu erfassen beabsichtigte: nämlich militärische Führungsleistung.

Unterschiede zwischen den Auswertungsstandards können anhand des TKML-Datensatzes gezeigt werden, aber nur, indem eine Gruppe mit einem sehr geringen Ausmaß an Expertise isoliert wird und ihre Mittelwerte mit jenen anderer Gruppen verglichen werden. Abbildung 8.3 zeigt genau diese Art der Unterscheidung zwischen den oberen und unteren 25% der Kadetten an der US-Militärakademie in West Point. Für die oberen 25% beträgt die Korrelation mit den Experten .95 und die Steigung der Regressionsgeraden 1.00. Für die unteren 25% beträgt die Korrelation mit den Experten jedoch .85, und die Höhe des Steigungskoeffizienten liegt nur noch .31. Die geringe Steigung weist auf eine eingeschränkte Varianz in den Itemmittelwerten hin, die auf Basis der untersten 25% der Kadetten berechnet wurden. Nur durch eine künstliche Einschränkung der Untersuchungsstichprobe auf das unterste Quartil der Kadettenstichprobe können substantielle Änderungen in den Standards bewirkt werden, und selbst dann beträgt die Korrelation immer noch .85.

Wenn unsere Vorstellungen darüber, wie Expertenwissen mit konsensbasierten Skalen erfasst wird, korrekt sind, sollten Anfänger nicht nur eine geringere Korrelation mit Experten aufweisen als Fortgeschrittene auf dazwischenliegenden Expertisestufen, sondern die Steigung der Regressionsgeraden sollte ebenfalls geringer sein. Um diese Vorhersage nachvollziehen zu können, bedenke man, auf welche Weisen die vielen verschiedenen und weniger richtigen Ansichten von Anfängern kombiniert werden können. Wenn keine systematischen Verzerrungen vorliegen, sollten einzelne Aspekte der Ansichten von Anfängern in verschiedener Hinsicht falsch sein, aber die Aspekte, die nahezu als Expertise zu bezeichnen sind, sollten ähnlich sein. Je mehr Fehler gemacht werden, desto stärker sollte die Regression zur Mitte ausfallen und folglich sollten gerin-

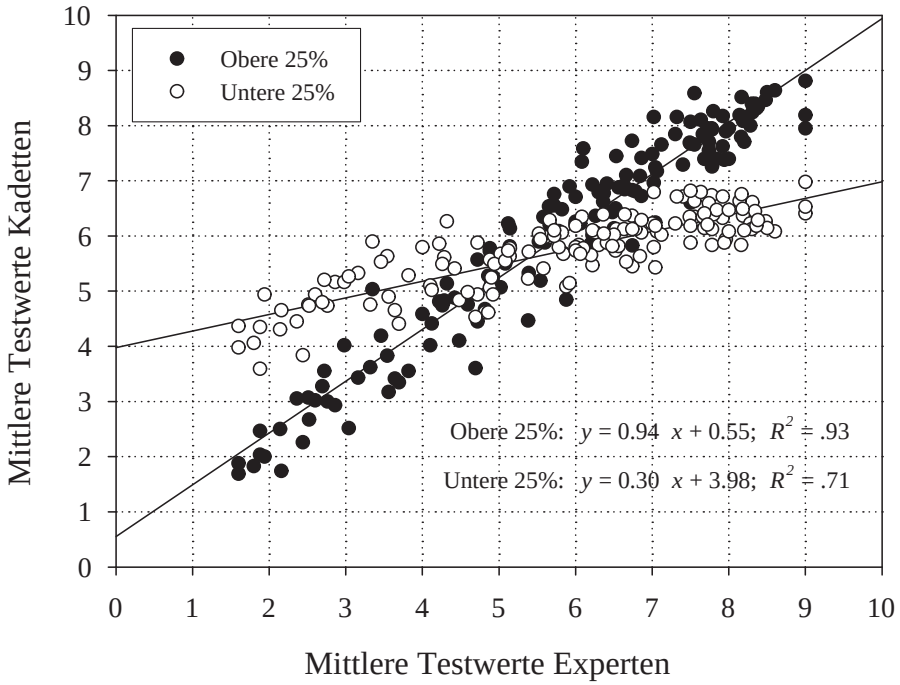

Abbildung 8.3 Zusammenhang zwischen den besten 25% aller Kadetten, den schlechtesten 25% und den als Experten für die Standardisierung des TKML herangezogenen Offizieren mit höherem Laufbahnstatus. Die Abbildung zeigt, dass die besten 25% von der Gruppe der Offiziere praktisch nicht zu unterscheiden ist.

gere Steigungen resultieren. Somit sind die TKML-Daten mit einem Modell konsistent, in dem sich Experten und Fortgeschrittene primär bezüglich ihrer Varianz unterscheiden und Veränderungen in der zentralen Tendenz stärker mit Unterschieden zwischen Anfängern und Fortgeschrittenen in Verbindung stehen. Dieses Modell wird in Abbildung 8.2d dargestellt.

8.6 Konzeptualisierung konsensbasierter Messung: auf dem Weg zu einem Arbeitsmodell

Zwei Ziele dieses Kapitels bestanden in der Beschreibung von KBM und der Zusammenfassung von Studien, die ihre Effektivität und Nützlichkeit ausweisen. Das anfängliche Modell war jedoch eher deskriptiv als theoretisch ausgerichtet und das Konzept, dass Expertenwissen durch die Heranziehung großer Mengen von Nicht-Experten angenähert werden kann, muss ein wenig eingeschränkt werden. Daher ist eine theoretische Erklärung von KBM von Interesse. Um konsensbasiertes Scoring zu verstehen, ist es nützlich sich vor Augen zu führen, dass sich das Wissen in den meisten Wissensbereichen, und insbesondere in Bereichen prozeduralen Wissens, als Folge von Erfahrungen ansammelt (siehe Anderson & Lebiere, 1998). Wenn eine größere Bandbreite an Ereignissen erlebt wird, werden höhere Stufen des Wissens und mit ihnen assoziierte Fertigkeiten erworben und die Reaktionen auf ein neues Ereignis oder eine neue Situation können ein zunehmendes Ausmaß des Entwicklungsstandes widerspiegeln.

Wenn Novizen in einem Wissensbereich eine dort angesiedelte Aufgabe zur Analyse vorgelegt wird, haben sie relativ gesehen wenige Grundlagen für ihre Ansichten und werden häufig weder untereinander noch mit Experten einer Meinung sein. Meinungsverschiedenheiten zwischen Novizen werden erwartet, weil das Wissen und die kognitiven Strukturen eines einzelnen Novizen entweder die Inhalte einiger weniger Erfahrungen oder die Inhalte von Erfahrungen, die nur eine geringfügige Bedeutung für die vorliegende Situation haben, widerspiegeln. Daher werden sich Novizen auf unterschiedliche Erfahrungen und Erwartungen beziehen und ihre Ansichten werden sowohl zu Widersprüchen untereinander als auch mit Experten führen.

Im Gegensatz dazu haben Experten in der Regel gut entwickelte und reife Wissensstrukturen, die einen breiten, umfassenden Erfahrungsschatz repräsentieren. Obwohl auch Experten in der Regel unterschiedliche Erfahrungshintergründe haben, überschneiden sich diese häufig zu großen Teilen. Mit zunehmendem Ausmaß an Expertise werden Wissensstrukturen und damit in Verbindung stehende Meinungen zunehmend konsistenter. Fortgeschrittene mit teilweise entwickelten und variierenden Ausmaßen an Expertise werden auf einem moderaten Niveau sowohl untereinander als auch mit Experten einer Meinung sein. Diese mäßige Übereinstimmung basiert auf der Entwicklung kognitiver Strukturen, die eine moderate, aber nicht umfassende Ansammlung von Erfahrungen widerspiegelt. Aus mathematischer Sicht kann die Korrelation des Wissens zwischen einer Person A und einer Person B als Produkt der Korrelation der Person A mit der „Wahrheit“ und der Person B mit der „Wahrheit“ aufgefasst werden. Wenn die Personen A und B mehr Wissen erwerben und ihre Ansichten „richtiger“ werden, korrelieren ihre Ansichten und Reaktionen höher (vgl. Romney & Weller, 1984).

Theoretisch betrachtet ist dieser Verlauf sinnvoll, aber in vielen Domänen sind Experten häufig nicht einer Meinung und ihre Leistung ist nicht sehr beeindruckend im Vergleich zur Leistung von Nicht-Experten: Die klinische Psychologie, Studienplatzvergabe und Vorhersagen in der Wirtschaft sind Beispiele für Bereiche, in denen Experten kaum bessere Leistungen zeigen als Laien (Chi, Glaser & Farr, 1988). Daher können Erwartungen an Expertenübereinstimmungen leicht überzogen sein. Eine realistischere Sichtweise ist die Erwartung, dass sich Experten quantitativ und nicht qualitativ von Fortgeschrittenen unterscheiden. Um den Experten gegenüber fair zu bleiben: Auch in diesen oben genannten Bereichen *können* sie bessere Leistungen zeigen als Anfänger.

Weil prozedurales Wissen auf Erfahrung beruht und diese Erfahrungen vom Auftreten von Ereignissen im Alltagsleben abhängig sind, können mehrere Fortgeschrittene unterschiedliche Arten von Erfahrungen und Wissen haben, obwohl ein Großteil dieses Wissens für häufig auftretende Situationen von großer Bedeutung ist. Daraus folgt, dass die Breite der Erfahrungen eines einzelnen Experten – obwohl sie umfassender ist als die eines einzelnen Fortgeschrittenen – häufig durch die Vielfalt der Erfahrungen von Fortgeschrittenen übertroffen wird. Die Bedeutung dieser Sichtweise für KBM ebenso wie für andere Anwendungen, in denen Wissen eine bedeutende Rolle spielt, ist, dass in den Wissensstrukturen einer großen Anzahl von Fortgeschrittenen mehr Informationen enthalten sein können als in denen einer kleinen Anzahl von Experten.

Das Konzept, dass Expertise die Gesamtsumme vieler kleiner Einzelbereiche darstellt, lässt sich insofern gut mit Intelligenztheorien in Verbindung bringen, als dass Intelligenz als eine allgemeine Lebensexpertise aufgefasst werden kann. Nach der Konzeptualisierung von Thomson (1928, 1939) entsteht der *g*-Faktor der Intelligenz durch die separate Wirkung vieler Verbindungen, deren Summe das Intelligenzniveau darstellt. Diese Auffassung basiert auf der Anwendung der Stichprobentheorie auf die Intelligenzmessung. Diesem Intelligenzmodell zufolge würde keine einzelne Person über

alle Verbindungen hinweg perfekte Leistungen zeigen, aber über alle Personen hinweg würden alle Verbindungen gelegentlich „geschlossen“ werden. Von IQ-Tests wurde angenommen, dass sie eine Stichprobe dieser Verbindungen zur Abschätzung des Gesamtniveaus der Konnektivität oder allgemeinen Intelligenz darstellen. Ein hoher IQ wiese eine hohe Anzahl an Verbindungen nach, ein geringer IQ eine niedrige Anzahl. Geringe und mäßige IQ-Werte könnten jedoch auch einfach von separaten und sich manchmal nicht überschneidenden Mengen von Verbindungen herrühren, beispielsweise wenn verschiedene Personen Faktenwissen in unterschiedlichen Bereichen besitzen, viele andere Fragen jedoch nicht beantworten können. Um im Rahmen moderner Intelligenzkonzepte zu bleiben kann *g* daher als die Summe einer Vielzahl separater Faktoren oder kognitiver Strukturen angesehen werden.

Weil Lerntheorien Wissen und Erfahrung miteinander in Verbindung setzen, ist Thomsons Sichtweise der Intelligenz als eine Repräsentation der Summe vieler kleiner Teile oder Verbindungen von Bedeutung. Expertise kann als die Reflexion der Gesamtzahl und -stärke der kognitiven Strukturen einer Person konzeptualisiert werden; genauso wie Intelligenz das Vorhandensein von Verbindungen widerspiegeln mag. Über verschiedene Personen hinweg kann gering ausgeprägte Expertise kognitive Strukturen widerspiegeln, die größtenteils sich nicht überschneidende Ereignismengen darstellen, wohingegen ein hohes Ausmaß an Expertise vollständigere Mengen kognitiver Strukturen und Erfahrungen darstellt. Wie in Thomsons Analysen kann von keiner einzelnen Person erwartet werden, dass sie die Gesamtheit aller mit einem bestimmten Bereich verbundenen Erfahrungen gemacht hat. Von einer großen Personenmenge, in der jede einzelne Person gewisse Erfahrungen besitzt, kann jedoch angenommen werden, dass die meisten Ereignisklassen – wenn nicht alle – besetzt sind und dass zu diesen Ereignissen korrespondierende kognitive Strukturen existieren.

Diese Lerntheorien sind hinsichtlich des Verständnisses von KBM von Bedeutung, wenn kognitive Strukturen und das mit ihnen verbundene Wissen die Erfahrung größtenteils unvorhersehbarer Ereignisse widerspiegeln (so wie es häufig beim prozeduralen und impliziten Wissen der Fall ist). Im Gegensatz dazu spiegelt akademisches Wissen eher formale Bildung wider, die oft so strukturiert ist, dass sie eine systematische, hochgradig geordnete und auf objektiven Informationen beruhende Erlebnismenge bereitstellt. Die Befragung von Studierenden zu noch nicht behandelten Themengebieten liefert mit großer Wahrscheinlichkeit nicht sehr viele Informationen. Die in diesem Kapitel beschriebenen Bereiche korrespondieren jedoch mit inzidentellem, implizitem oder prozeduralem Wissen. Was die Methodologie der SJTs (vgl. z. B. McDaniel et al., 2001) betrifft und – wie wir vermuten – viele „weiche“, unscharf definierte Domänen psychologischer Forschung, herrschen ähnliche Bedingungen vor.

Auf diese Weise unterstützen kognitive Theorien, die sich auf den Erwerb prozeduralen Wissens beziehen, die Behauptung, dass die Ansichten einer großen Anzahl von Fortgeschrittenen als Annäherung an die Ansichten einer kleinen Anzahl von Experten in diesen Bereichen verwendet werden können. Diese Ansicht stellt den Kern der KBM dar. In diesem Kapitel beschäftigten wir uns mit dem Einsatz von KBM für szenariobasierte Testverfahren mit Likert-Antwortformaten. Es sollte angemerkt werden, dass unsere Ergebnisse mit Simulationen konsistent sind, in denen dichotome Items verwendet werden, aber die objektiv richtigen Antworten nicht spezifiziert werden (Batchelder & Romney, 1988). Diese Analysen zeigen, dass äußerst genaue Antwortschlüssel konstruiert werden können, indem relativ kleine Probandengruppen verwendet werden, in denen die Anzahl der Probanden mit der Expertise der Gruppe im Gleichgewicht steht. Diese Daten zeigen auch, dass eine Mehrheitsregel angewendet werden kann, um bei

einer großen Probandenanzahl die richtigen Antworten zu bestimmen. Natürlich wird ein solches Vorgehen nur selten bei einem für einen geläufigen Wissensbereich entwickelten Test mit dichotomem Antwortformat benötigt, doch stehen diese Ergebnisse in Einklang mit unseren Befunden und Schlussfolgerungen, dass KBM sehr gut für unscharf definierte und im Entstehen begriffene Wissensbereiche geeignet ist. Es scheint uns wahrscheinlich, dass dieser Ansatz für die Testentwicklung in neuen Wissensbereichen so lange relevant bleiben wird, bis sie wesentlich genauer eingegrenzt werden. Dies gilt insbesondere für Bereiche, in denen Erfahrung eine große Rolle spielt, also auch für die emotionale Intelligenz.

8.7 Konsensbasierte Messung: Implikationen und Einschränkungen

Konsensbasiertes Scoring hat einige bedeutende Implikationen für die Untersuchung interindividueller Unterschiede. Erstens erlaubt dieser Ansatz die Konstruktion und das Scoring von Testverfahren für Wissensbereiche, für die es keine Experten gibt oder auf einfache Weise gefunden werden können. Dies erlaubt eine Ausweitung der Anwendungsbereiche, für die Wissenstests entwickelt werden können. Diese Ausweitung geht über traditionelle, formale Bereiche hinaus und reicht hinein in Wissensbereiche, die für unser alltägliches Leben bedeutungsvoll sind. Auf diese Weise erlaubt konsensbasiertes Scoring die Erfassung von Wissen in Bereichen, die in der psychologischen und pädagogischen Forschung traditionell bisher nicht untersucht wurden. Die Einsatzbereiche psychologischer Erhebungsverfahren und der Intelligenzforschung werden erweitert. Der Intelligenzbereich wird potenziell gleich um mehrere Aspekte erweitert; nicht zuletzt könnten sich darunter emotionale oder soziale Intelligenz befinden. Diese Auffassung ist mit Theorien impliziten Wissenerwerbs vereinbar und lässt sich gut mit Konzeptualisierungen sozialen Wissens in Verbindung bringen.

Eine zweite bedeutsame Implikation von KBM ist, dass sie eine ökonomische Testentwicklung ermöglicht. Der Ansatz erlaubt es, Fragen zu stellen, zu beantworten und zu bewerten, ohne dass die korrekten Antworten im Vorhinein bekannt sein müssen. Folglich wird der Entwicklungszyklus für Testverfahren verkürzt, da keine Expertenantworten zur Konstruktion von Auswertungsstandards benötigt werden. Zusätzlich werden die Kosten, die mit der Erstellung von Auswertungsstandards und -rubriken verbunden sind, minimiert, denn die Einholung von Expertenurteilen kann sehr teuer sein, wohingegen Probandendaten bei der Skalenerstellung unmittelbar anfallen. Eine ähnliche Implikation zur Unterstützung von KBM resultiert aus der Verwendung von Likert-Skalen. Diese erlauben die Berechnung von Unterschieden auf der Itemebene und folglich eine vollständigere Analyse der erhaltenen Informationen. Wie bei der Verwendung zusätzlicher Informationen erwartet werden kann, weisen Vergleiche von Testwerten, die auf Distanzinformation beruhen, höhere Reliabilitäten auf als solche, die auf einem dichotomen Antwortformat basieren (Legree, 1995). Deshalb unterstützt das Likert-Format eine verbesserte Testeffizienz. Zusätzlich können Distanzitems korreliert werden, und aus diesen Korrelationen ließen sich inhaltlich interpretierbare Faktoren extrahieren (siehe Legree et al., 2003).

Drittens erlaubt KBM die Auswertung desselben Protokolls mit Hilfe multipler Standards. Dieser Ansatz könnte bei der Untersuchung uneindeutiger Wissensbereiche, für die verschiedene Gruppen unterschiedliche Ansichten vertreten, nützlich sein. Er könnte gut zum Verständnis differierender Ansichten bezüglich Geschlechtern, Personen ver-

schiedener politischer Zugehörigkeit, Rassenzugehörigkeit, sexueller Orientierung oder unterschiedlichen Alters eingesetzt werden oder bei der Identifizierung der Grundlagen konkurrierender Theorien, die ein bestimmtes Phänomen erklären, hilfreich sein. Der Ansatz könnte sogar auf die Entwicklung formaler Theorien der Skalierung und Itemmessung angewendet werden, beispielsweise auf konsensuelle Skalierung, und zwar in einem rekursiven Zyklus!

Viertens beinhaltet KBM explizit die Annahme von Unstimmigkeiten und Widersprüchen in der Kohärenz von Wissensstrukturen. Unscharf definierte Bereiche werden dadurch charakterisiert, dass selbst unter Experten Unstimmigkeiten existieren. Faktorenanalysen und multidimensionale Skalierungen ihrer Antworten (Psotka et al., 2004), die so leistungsfähige Techniken wie die latente semantische Analyse verwenden, bringen nicht nur Ordnung in diese Unstimmigkeiten, sondern geben auch Anlass zu der Hoffnung, die Quelle der Differenzen definieren zu können und neue Verbindungen innerhalb der informellen Rahmenkonzepte zu schaffen. Um eine wohl bekannte Ansicht zu paraphrasieren:[3] „An intuitive inconsistency is the muse of great minds".

Fünftens betont konsensbasiertes Scoring, dass *zumindest unter gewissen Bedingungen* die Standards, die auf einer Gruppe spezifisch kenntnisreicher Personen basieren, den Standards von Experten nahe kommen. SJTs werden gelegentlich als wenig vertrauenswürdig bezeichnet, weil sie bereits Stimulusreize zur Situationsbeschreibung mit ausreichender Spezifität bieten, um Antworten hervorzurufen, die das zur Messung angestrebte Phänomen repräsentieren. Unsere Hauptinterpretation legt nahe, dass Urteile über diese mehrdeutigen Situationen das existierende Wissen direkt widerspiegeln. Eine von der Gestaltpsychologie inspirierte alternative Erklärung ist, dass abstrakte Stimulussituationen nicht alle für eine Antwort benötigten Reize bieten und stattdessen eine Interpretation oder Induktion der Bedeutung erzwingen. Somit reflektieren die Antworten eher die über das Verständnis der abstrakten Situationen vermittelten existierenden Wissensstrukturen als direkt die Qualität existierender Strukturen. Sehr gute Leistungen würden dann einen besseren Zugang zu einer gemeinsamen Vorstellungswelt reflektieren, indem Interpretationen und Induktion von Bedeutungen erzwungen werden.

Unter den folgenden Bedingungen scheint es weniger wahrscheinlich zu sein, dass KBM eine nützliche Metrik der Gruppenübereinstimmung hervorbringen wird, die zur Bewertung von Expertise notwendig ist: Paradigmenwechsel in der Forschung oder wenn Informationen so verteilt werden, dass Ansichten von Experten oder Fortgeschrittenen differenziell beeinflusst werden, oder auch wenn diese Bedingungen zu Gruppenunterteilungen führen, die den Gruppenzielen eher entgegenstehen als ihnen zuträglich zu sein. Ob ein multi-modaler Ansatz zur Entwicklung multipler Metriken verwendet werden kann, ist eine ungeklärte Frage. Aber dieser Ansatz könnte für das Verständnis von Interaktionen zwischen Gruppen, die sich teilweise widersprechen, von Bedeutung sein.

Soziales Wissen repräsentiert oftmals die Konvergenz zwischen verschiedenen Perspektiven, und es wird allgemein angenommen, dass am Schnittpunkt dieser Perspektiven die Wahrheit liegt. Folglich stellt das amerikanische Rechtssystem, in dem eine Seite als Ankläger und die Gegenseite als Verteidiger fungiert, eine Manifestierung dieser Sichtweise dar, ebenso wie alle demokratischen Institutionen. Die Sichtweise, dass Wissen in vielen verschiedenartigen Ansichten verwurzelt ist, spiegelt sich in Tolstojs Beobachtung „Glückliche Familien sind alle gleich; jede unglückliche Familie ist auf ihre eigene Weise unglücklich" und aus einer kulturübergreifenden Sichtweise in dem

[3] „A foolish consistency is the hobgoblin of small minds" – Ralph Waldo Emerson.

afrikanischen Sprichwort „Es braucht ein Dorf, um ein Kind großzuziehen“ wider. Der Erfolg dieser Institutionen und die Relevanz dieser Aussagen zeigen sich in der Vorstellung, dass Wissen über mehrere Personen verteilt sein kann. Um dieses Wissen optimal ausschöpfen zu können, bieten sich konsensbasierte Messmethoden zur Analyse seiner Struktur und seiner unzweifelhaft vorhandenen Nützlichkeit in jungen Disziplinen wie der emotionalen und sozialen Intelligenz geradezu an.

8.8 Epilog

Von unseren Lesern hätten wir gerne eine Rückmeldung. Bitte schicken Sie uns eine E-Mail mit Ihren Ratings – unter Verwendung einer 9-stufigen Likert-Skala – zu welchem Ausmaß (1 = überhaupt nicht ... 9 = absolut) Sie glauben, dass:

1. Sie sich gut mit Testentwicklung auskennen.
2. traditionelle Methoden der Testentwicklung für gut spezifizierte Wissensdomänen geeignet sind.
3. traditionelle Methoden der Testentwicklung für im Entstehen begriffene, unscharf abgegrenzte Wissensbereiche geeignet sind.
4. konsensbasierte Messmethoden für gut spezifizierte Wissensbereiche geeignet sind.
5. konsensbasierte Messmethoden für im Entstehen begriffene, unscharf abgegrenzte Wissensbereiche geeignet sind.
6. akademisches Wissen mit Multiple-choice-Items genau erfasst werden kann.
7. akademisches Wissen mit Likert-Items genau erfasst werden kann.
8. prozedurales Wissen mit Multiple-choice-Items genau erfasst werden kann.
9. prozedurales Wissen mit Likert-Items genau erfasst werden kann.
10. es richtig ist anzunehmen, dass glückliche Familien einander ähnlicher sind als unglückliche Familien.

Wenn wir genügend Rückmeldungen erhalten, werden wir die Antwortverteilungen für diese zehn Items von Lesern dieses Buchs mit denen von Testentwicklern vergleichen, die diesen Überblick nicht gelesen haben. Wenn unsere Theorie richtig ist, sollte sich ein höheres Maß an Übereinstimmung für Items mit KBM unter den Lesern dieses Kapitels zeigen als unter den Nicht-Lesern, was durch geringere Varianz bei ähnlichen Mittelwerten der Items belegt würde. Bitte schreiben Sie eine E-Mail an den ersten Autor unter der Adresse legree@ari.armi.mil.

Anmerkungen der Autoren

Die Ansichten, Meinungen und/oder Befunde, die in diesem Artikel enthalten sind, sind einzig die der Autoren und sollten nicht als offizielle Position des *Department of the Army* oder *DOD* oder als Grundsatz oder Entscheidung ausgelegt werden, außer wenn sie durch anderweitige Dokumentation entsprechend gekennzeichnet sind.

Literatur

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum.

Arthur, J., W., Barrett, G. V. & Alexander, R. A. (1991). Prediction of vehicular accident analysis: A meta–analysis. *Human Performance*, *4*, 89–105.

Batchelder, W. H. & Romney, A. K. (1988). Test theory without an answer key. *Psychometrica*, *53*, 71–92.

Cantor, N. & Kihlstrom, J. F. (1987). *Personality and social intelligence*. Englewood Cliffs, NJ: Prentice-Hall.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Chi, M. T. H., Glaser, R. & Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum.

Detterman, D. K., Caruso, D. R., Mayer, J. D., Legree, P. J. & Conners, F. (1992). Assessment of basic cognitive abilities in relation to cognitive deficits; mopping up: Relation between cognitive processes and intelligence. *American Journal on Mental Retardation*, *97*, 251–286.

Hedlund, J., Forsythe, G. B., Horvath, J. A., Williams, W. M., Snook, S. & Sternberg, R. J. (2003). Identifying and assessing tacit knowledge: Understanding the practical intelligence of military leaders. *Leadership Quarterly*, *14*, 117–140.

Heffner, T. S. & Porr, W. B. (2000, August). *Scoring situational judgment tests: A comparison of multiple standards using scenario response alternatives.* Paper presented at the Annual Conference of the American Psychological Association, Washington, DC.

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods*, *1*, 131–163.

Jensen, A. R. (1998). *The g factor*. Westport, CT: Praeger.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor. *Intelligence*, *21*, 247–266.

Legree, P. J. & Grafton, F. C. (1995). *Evidence for an interpersonal knowledge factor: The reliability and factor structure of tests of interpersonal knowledge and general cognitive ability* (ARI Technical Report No. 1030). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Legree, P. J., Heffner, T. S., Psotka, J., Martin, D. E. & Medsker, G. J. (2003). Traffic crash involvement: Experiential driving knowledge and stressful contextual antecedents. *Journal of Applied Psychology*, *88*, 15–26.

Legree, P. J., Martin, D. E. & Psotka, J. (2000). Measuring cognitive aptitude using unobtrusive knowledge tests: A new survey technology. *Intelligence*, *28*, 291–308.

Mayer, J. D., Caruso, D. R. & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, *27*, 267–298.

Mayer, J. D., Salovey, P., Caruso, D. R. & Sitarenios, G. (2003). Modeling and measuring emotional intelligence with the MSCEIT V2.0. *Emotion*, *3*, 97–105.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A. & Braverman, E. P. (2001). Use of situational judgement tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730–740.

Psotka, J. (1977). Syntely: Paradigm for an inductive psychology of memory, perception, and thinking. *Memory and Cognition*, *3*, 553–600.

Psotka, J., Streeter, L. A., Landauer, T. K., Lochbaum, K. E. & Robinson, K. (2004). *Augmenting electronic environments for leadership.* In Advanced Technologies for Military Training: Proceedings No. RTO-MP-HFM-101-21 of the Human Factors and Medicine Panel, Genoa, Italy, October 13, 2003 (pp. 287–301). Neuilly-sur-Seine, France: Research and Technology Organization.

Roberts, R. D., Zeidner, M. & Matthews, G. (2001). Does emotional intelligence meet traditional standards for an intelligence? Some new data and conclusions. *Emotion*, *1*, 196–231.

Romney, A. K. & Weller, S. C. (1984). Predicting informant accuracy from patterns of recall among informants. *Social Networks*, *6*, 59–77.

Schaie, K. W. (2001). Emotional intelligence: Psychometric status and developmental characteristics—Comment on Roberts, Zeidner, and Matthews. *Emotion*, *1*, 243–248.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M. et al. (2000). *Practical intelligence in everyday life.* New York: Cambridge University Press.

Thomson, G. H. (1928). A worked out example of the possible linkages of four correlated variables on the sampling theory. *The British Journal of Psychology*, *18*, 68–76.

Thomson, G. H. (1939). *The factorial analysis of human ability.* New York: Houghton-Mifflin Company.

Veling, I. H. (1982). Measuring driving knowledge. *Accident, Analysis, & Prevention*, *14*, 81–85.

Zeidner, M., Matthews, G. & Roberts, R. D. (2001). Slow down you move too fast: Emotional intelligence remains an elusive intelligence. *Emotion*, *1*, 265–275.