Bundesministerium
für Bildung
und Forschung

20 Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik

Bildungsforschung Band 20

Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik

Impressum

Herausgeber
Bundesministerium
für Bildung und Forschung (BMBF)
Referat Öffentlichkeitsarbeit
11055 Berlin

Bestellungen
Schriftlich an den Herausgeber
Postfach 30 02 35
53182 Bonn

oder per
Tel.: 01805-262 302
Fax: 01805-262 303
(0,14 Euro/Min.)

E-Mail: books@bmbf.bund.de
Internet: http://www.bmbf.de

Autoren
Dr. Johannes Hartig (Deutsches Institut für Internationale Pädagogische Forschung), Frankfurt am Main
Prof. Dr. Eckhard Klieme (Deutsches Institut für Internationale Pädagogische Forschung), Frankfurt am Main
Nina Jude (Deutsches Institut für Internationale Pädagogische Forschung), Frankfurt am Main
Astrid Jurecka (Deutsches Institut für Internationale Pädagogische Forschung), Frankfurt am Main
Ulf Kröhne (Deutsches Institut für Internationale Pädagogische Forschung), Frankfurt am Main
Prof. Dr. Katharina Maag-Merki (Deutsches Institut für Internationale Pädagogische Forschung), Frankfurt am Main
Dr. Jean-Paul Reef (Deutsches Institut für Internationale Pädagogische Forschung), Frankfurt am Main
Dr. Joachim Wirt (Deutsches Institut für Internationale Pädagogische Forschung), Frankfurt am Main

Bonn, Berlin 2007

Gedruckt auf Recyclingpapier

Johannes Hartig
Eckhard Klieme
(Hrsg.)

Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik

Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung

Inhalt

1 Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen

Eckhard Klieme, Katharina Maag-Merki & Johannes Hartig

Infolge der zunehmenden Wissensintensität in vielen Arbeits- und Lebensbereichen und der Globalisierung von Arbeits- und Bildungsmärkten wird die Frage nach der Produktivität des Bildungswesens zu einer gesellschaftlichen Kernfrage. Von der Bildungsforschung wird erwartet, dass sie diese Produktivität messbar macht, Erklärungsmodelle für Verlauf, Effektivität, und Effizienz von Bildungsprozessen bereitstellt und Interventionsstrategien wissenschaftlich untersucht. Diese Anforderungen wachsen in dem Maße, in dem das Bildungswesen selbst zum Gegenstand internationalen Wettbewerbs wird, beispielsweise durch die Vergleichsstudien der OECD zu life skills von Jugendlichen am Ende der Pflichtschulzeit (PISA) oder durch die einheitliche Regelung von Studienzulassung, Studienverlauf und Zertifizierungen im Rahmen des Bologna-Prozesses. Um zu beschreiben, inwieweit Individuen den Anforderungen in verschiedenen Kontexten gewachsen sind, wird häufig der Begriff der *Kompetenz* verwendet. Über Kompetenz in einem bestimmten Bereich zu verfügen, bedeutet in diesem Bereich erfolgreich handeln zu können; Inkompetenz heißt, den Anforderungen in einem Bereich nicht gewachsen zu sein.

Kompetenz hat bereits in der Alltagssprache sehr vielfältige Bedeutungen, z.B. auch die Zuständigkeit im juristischen Sinne. Die für die Bildungsforschung relevante alltagssprachliche Bedeutung ist Kompetenz als Synonym zu *Fähigkeit* (Duden, 2001). Wie häufig bei der Übernahme alltagssprachlicher Begriffe in wissenschaftliche Kontexte ist auch der Begriff der Kompetenz in der Bildungsforschung unscharf geblieben; in unterschiedlichen Fachdisziplinen und Forschungsfeldern werden spezifische, zum Teil erheblich voneinander abweichende Kompetenzbegriffe verwendet. Auch ein nur halbwegs vollständiger Überblick über die Bedeutungen von Kompetenz in unterschiedlichen Bereichen der nationalen und internationalen Bildungsforschung der letzten Jahrzehnte erscheint angesichts der Vielfalt und Menge von theoretischen und empirischen Arbeiten ein enorm aufwendiges Unterfangen. Eine Stichwortsuche in der Literaturdatenbank des Fachinformationssystems Bildung (FIS Bildung) des Deutschen Instituts für Internationale Pädagogische Forschung (DIPF) liefert für *Kompetenz* 8.889 Treffer, in der Datenbank PsycInfo finden sich ab 1985 für *competence*, *competency* und *competencies* 27.255 Treffer – das entspricht drei bis vier Veröffentlichungen pro Tag in diesem Zeitraum. Es erscheint daher unerlässlich, für einen konkreten Forschungskontext eine schärfende Eingrenzung des Kompetenzbegriffs vorzunehmen. In diesem Kapitel soll skizziert werden, wie der Begriff *Kompetenz* im Kontext der vorliegenden Expertise verstanden werden soll. Bei dieser Begriffspräzisierung ist vor allem von Interesse, welche Eigenschaften einer Kompetenzdefinition nützlich sind, wenn der Begriff im Kontext der pädagogisch-psychologischen Diagnostik verwendet werden soll und „Kompetenz" zur Charakterisierung der *Ergebnisse von Bildungsprozessen* verwendet werden soll.

1.1 Hintergrund und Eingrenzung des Kompetenzbegriffs

Obwohl schon vor Jahrzehnten eingeführt, ist der Kompetenzbegriff erst in den letzten Jahren zum Gegenstand intensiver Diskussion in der Psychologie und ihren Nachbardisziplinen geworden (z.B. Csapó, 2004; Klieme, Funke, Leutner, Reimann & Wirth, 2001; Sternberg & Grigorenko, 2003; Rychen & Salganik, 2001, 2003; Weinert, 2001b). Nicht zuletzt greift die Forschung diesen Begriff auf, um veränderte Anforderungen der Lebens- und Arbeitswelt sowie damit zusammenhängende Bildungsziele zu charakterisieren.

In der pädagogisch-psychologischen Diagnostik wurde der Kompetenzbegriff als Gegenbegriff zu generalisierten, kontextunabhängigen kognitiven Leistungskonstrukten eingeführt, wie sie für die Intelligenzforschung und -diagnostik typisch sind. In einer pointierten Kritik an diesen Konstrukten („Testing for competence rather than for ‚intelligence'") wird die Erfassung von „Kompetenzen" von McClelland (1973) mit der Hoffnung verbunden, eine bessere Passung zwischen Testinhalten und Anforderungen in realen (z.B. beruflichen) Situationen und damit eine bessere Vorhersage von Leistungsunterschieden in diesen Situationen zu erzielen. Empirische Unterstützung für die von McClelland geübte Kritik an der traditionellen kognitiven Leistungsdiagnostik steht allerdings aus (z.B. Barrett & Depinet, 1991). Inhaltlich bezeichnen „Kompetenzen" für McClelland (1973) die für eine spezifische Tätigkeit notwendigen Voraussetzungen, wobei er selbst keine genauere konzeptuelle oder theoretische Klärung des Begriffs vornimmt. Aus dieser Perspektive kann jedes beliebige Konstrukt als „Kompetenz" betrachtet werden, wenn es der Vorhersage der Bewährung in konkreten Leistungssituationen dient: „Some of these competencies may be rather traditional cognitive ones involving reading, writing, and calculating skills. Others should involve what traditionally have been personality variables, although they might better be considered competencies." (S. 10).

Ein Schlüsselmerkmal des Kompetenzbegriffs in der psychologisch-pädagogischen Diagnostik ist fachhistorisch also zunächst der stärkere Bezug zum „wirklichen Leben", z.B. in Gestalt von Anforderungen in beruflichen Kontexten. Dieser Bezug zu konkreten Kontexten findet sich durchgehend in verschiedenen Definitionen von Kompetenz. So definierte bereits White (1959, S. 317) Kompetenz als „effective interaction (of the individual) with the environment"; besonders prägnant bezeichnen Connell, Scheridan und Gardner (2003, S. 142) Kompetenzen als „realized abilities". Auch für Weinert (1999, 2001), der in einem für die OECD erstellten Gutachten eine Übersicht über verschiedenen Kompetenz-Definitionen vornahm, ist die *Kontextspezifität* von Kompetenzen zentral (vgl. auch Klieme, 2004a). Während in der Intelligenzforschung kognitive Leistungskonstrukte untersucht werden, die über eine breite Vielfalt von Situationen generalisierbar sind, beziehen sich Kompetenzkonstrukte auf spezifische Anforderungsbereiche – die Frage *„kompetent wofür?"* ist notwendiger Bestandteil jeder Kompetenzdefinition.

Weinert (1999, 2001a) empfiehlt unter Abwägung unterschiedlicher theoretischer und pragmatischer Argumente darüber hinaus, Kompetenzen als kontextspezifische *kognitive* Leistungsdispositionen definieren. Diese spezifischen Leistungsdispositionen lassen sich auch als Kenntnisse, Fertigkeiten oder Routinen charakterisieren. Zusätzlich zur Kontextspezifität wird hiermit noch eine

weitere Einschränkung des Kompetenzbegriffs vorgenommen, nämlich auf kognitive Leistungsdispositionen; motivationale oder affektive Voraussetzungen für erfolgreiches Handeln werden damit ausgeschlossen. Diese Einschränkung ist nicht selbstverständlich; Weinert selbst diskutiert auch so genannte Handlungskompetenzen, die motivationale Orientierungen, Einstellungen, Tendenzen und Erwartungen einschließen. Allerdings schlägt Weinert an derselben Stelle vor, in empirischen Untersuchungsdesigns kognitive und motivationale Tendenzen getrennt zu erfassen, weil nur so ihre Zusammenhänge mit kognitiven Voraussetzungen für erfolgreiches Handeln der empirischen Untersuchung zugänglich werden.

Aus der Kontextabhängigkeit von Kompetenzen ergibt sich ein weiterer wichtiger Aspekt des Kompetenzkonzepts: dass Kompetenzen durch Lernen erworben werden können bzw. erworben werden müssen (vgl. Hartig & Klieme, 2006). Der Bezug der Kompetenzdefinition auf spezifische Situationen und Anforderungen legt nahe, dass der Kompetenzerwerb das Sammeln von Erfahrungen in den entsprechenden Situationen bzw. mit den entsprechenden Aufgaben voraussetzt. Eine wie die oben verwendete Definition impliziert, dass Kompetenzen erworben sowie durch äußere Interventionen beeinflusst werden können (z.B. Baumert, Stanat & Demmrich, 2001; Hartig & Klieme, 2006; Simonton, 2003) und stellt damit eine weitere Abgrenzung zu kognitiven Grundfunktionen dar, die in wesentlich geringerem Maße erlernbar und trainierbar sind (Weinert, 2001).

Zusammenfassend werden Kompetenzen im Kontext der vorliegenden Expertise also als *kontextspezifische kognitive Leistungsdispositionen*, die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen, definiert (Klieme & Leutner, 2006). Diese Verwendung des Kompetenzbegriffs deckt sich auch mit der in den großen internationalen Schulleistungsstudien (PISA, TIMSS, PIRLS). Mit dieser Arbeitsdefinition werden zwei wesentliche Restriktionen vorgenommen: Zum einen sind Kompetenzen funktional bestimmt und somit bereichspezifisch auf einen begrenzten Sektor von Kontexten bezogen. Zum anderen wird die Bedeutung des Begriffs auf den kognitiven Bereich eingeschränkt, motivationale oder affektive Voraussetzungen für erfolgreiches Handeln werden explizit nicht mit einbezogen.

Die hier verwendete Definition von Kompetenzen ist in mehrfacher Hinsicht nützlich, wenn Ergebnisse von Bildungsprozessen beschrieben werden sollen:

- Der Bezug auf spezifische Kontexte gewährleistet eine hinreichende konzeptuelle Abgrenzung von allgemeinen kognitiven Leistungskonstrukten, die in der Literatur als nur in relativ geringem Umfang förderbar betrachtet werden.
- Der Bezug auf spezifische Kontexte erlaubt eine Definition spezifischer Kompetenzen, die jeweils an die Ziele spezifischer Bildungsmaßnahmen oder -einrichtungen angepasst werden kann. Hierdurch kann gewährleistet werden, dass Kriterien zur Evaluation in *Passung zu den intendierten Zielen der zu evaluierenden* Maßnahmen definiert werden.
- Die Begrenzung auf kognitive Dispositionen bedeutet, dass motivationale und affektive Variablen im Bildungsgeschehen separat erfasst werden und Zusammenhänge zwischen Kompetenzen und diesen nicht-kognitiven Vari-

ablen einer expliziten empirischen Untersuchung zugänglich sind. Auch allgemeine kognitive Fähigkeiten werden von Kompetenzen konzeptuell abgegrenzt und können entsprechend getrennt erfasst werden. Diese konzeptuelle Abgrenzung und separate Messung von möglichen interindividuell variierenden Einflussgrößen schafft die Grundlage, Effekte unterschiedlicher *individueller Ausgangsbedingungen* sowie *Entwicklungsprozesse* im Bildungsgeschehen differenzierter zu untersuchen.

Trotz der genannten Vorteile der hier verwendeten Arbeitsdefinition von Kompetenzen muss jedoch auch darauf hingewiesen werden, dass der Definition von „Kontext" eine kritische Bedeutung für die konkrete Definition einzelner Kompetenzen zukommt. Mit Kontext ist gemeint, auf welchen Bereich von Situationen und Anforderungen sich ein spezifisches Kompetenzkonstrukt bezieht. Diese Definition stellt eine für jede Fragestellung separat zu lösende Aufgabe und eine unter Umständen durchaus schwierige Herausforderung dar. Tatsächlich birgt die Definition von Kompetenzen als kontextspezifische kognitive Leistungsdispositionen die Möglichkeit einer gewissen Willkür, wenn der „Kontext" zu beliebig definiert wird.

Der relevante Kontext für die Definition eines Kompetenzkonstrukts muss einerseits hinreichend konkret sein, sollte andererseits auch nicht zu eng gefasst sein, da sonst einfaches Sachwissen oder isolierte Fertigkeiten unnötigerweise als Kompetenzen etikettiert werden. Ein nahe liegendes Kriterium, um die mögliche Willkür von Kontext- und damit Kompetenzdefinitionen einzuschränken, ist der Bezug auf eine Menge hinreichend *ähnlicher realer Situationen*, in denen bestimmte, ähnliche Anforderungen bewältigt werden müssen. „Real" könnte hierbei pragmatisch mit „außerhalb des Bildungsprozesses" übersetzt werden. So könnte etwa „Fremdsprachkompetenz" als die Fähigkeit zum erfolgreichen mündlichen und schriftlichen Kommunizieren in der jeweiligen Sprache definiert werden. Die Menge hiermit eingeschlossener Situationen ist hierbei schon recht groß, dennoch lassen sich gemeinsame ähnliche Anforderungen beschreiben. Wortschatz, die Kenntnis grammatischer Regeln und das Beherrschen der Ausspracheregeln wären bei einer derartigen Definition Voraussetzungen für Fremdsprachkompetenz. Das Beispiel illustriert auch, dass mit Hilfe des Konzepts der Kompetenz eine mögliche Unterscheidung zwischen reinem Wissen (z.B. über grammatische Regeln) und dessen Anwendung (z.B. in einer kommunikativen Situation) vorgenommen werden kann.

1.2 Kompetenzmessung zur Steuerung von Bildungssystemen

Die weltweit seit Ende der 1980er Jahre zu beobachtende Durchsetzung neuer Steuerungsstrategien für staatliches Handeln führt auf allen Stufen des Bildungswesens – von der Elementarbildung über Schulen und Hochschulen bis zur beruflichen Weiterbildung und Erwachsenenbildung – zu einem stärkeren Blick auf die Ergebnisse von Bildungsprozessen, die mit Begriffen wie *Output* oder *Outcome* bezeichnet werden. Die Produktivität von Bildungssystemen, die Qualität einzelner Bildungseinrichtungen und der Lernerfolg von Individuen soll messbar gemacht werden, um Bildungsprozesse wirksamer lenken zu können. Das bedeutet, dass dem Setzen und Überprüfen von *Zielen* und *Erwartun-*

gen, die Schulen zu erreichen bzw. zu erfüllen haben, eine zunehmende Bedeutung für die Qualitätsentwicklung des Bildungssystems gegenüber inhaltlichen Vorgaben über den *Input* wie zum Beispiel Lehrmitteln oder Lehrplänen eingeräumt wird. Entsprechende Erhebungs-, Auswertungs- und Feedbackverfahren, die wissenschaftlichen Ansprüchen genügen können, fehlen jedoch vielfach. Die neuen Steuerungsstrategien und insbesondere die Wirkungen und Nebenwirkungen eines verstärkten Einsatzes von Tests – z.B. das mögliche „teaching to the test" – sind durchaus umstritten (Fuhrman & Elmore, 2004; Hamilton, 2003; Klieme, 2004b). Umso mehr kommt es darauf an, die Qualität und Aussagekraft der neuen Messverfahren für pädagogische Handlungsfelder durch Grundlagenforschung zu untermauern und ggfs. kritische Entwicklungen aufzudecken.

Der Kompetenzbegriff ist für empirische Studien, die sich mit der Produktivität des Bildungswesens befassen, zentral geworden. Traditionelle Ansätze der pädagogisch-psychologischen Diagnostik – wie etwa die kriteriumsorientierte Leistungsmessung der 1970er Jahre, die hierarchisch gestufte, letztlich fachinhaltlich bestimmte Ziele in Testaufgaben umsetzte – stoßen heute an ihre Grenzen, weil sich die Bildungsziele selbst geändert haben (Segers, Dochy & Cascallar, 2003). Bildung und Qualifizierung lassen sich in einer modernen Industriegesellschaft nicht mehr durch einen festen Kanon fachlicher Kenntnisse, die an die nachfolgende Generation weiterzugeben sind, beschreiben. Wissen muss auf unterschiedliche, auch neue und komplexe Situationen und Kontexte anwendbar sein. Was Menschen gelernt haben, muss anschlussfähig sein für eigenständiges Weiterlernen. Neben dem Erwerb von vernetztem, anwendungsfähigem Wissen werden auch die Fähigkeit zum selbstregulierten Handeln, insbesondere zum selbstständigen Lernen, Problemlösefähigkeiten sowie soziale und kommunikative Fähigkeiten zu neuen Bildungs- und Qualifizierungszielen. Der Begriff der Kompetenz ist, wie eingangs erwähnt, mit dem Ziel verbunden, eben solche komplexen und realitätsnäheren Konstrukte fassbar zu machen. Kompetenzen wurden zunächst in der Weiterbildung und der beruflichen Bildung, später auch in der allgemeinen Schulbildung und der Hochschulbildung zum zentralen Konzept zur Charakterisierung der Zielsetzungen von Bildungssystemen.

Ein theoretisch und praktisch wichtiger Beitrag der neueren Bildungsforschung besteht daher in der Re-Konzeptualisierung und Operationalisierung von Bildungszielen unter dem Leitbild von Kompetenz sowie verwandter Konzepte wie *literacy* und *life skills*. Hierbei werden traditionelle Diskurse der Erziehungswissenschaft und der Psychologie über Bildung und Qualifikation, fachliches und fächerübergreifendes Lernen, Schlüssel- und Basisqualifikationen aufgegriffen und einer empirischen Untersuchung zugänglich gemacht.

Mit dem Vorschlag, Kompetenzmodelle zur Grundlage von Bildungsstandards zu machen (Klieme et al., 2003), hat die Forschung in diesem Bereich auch in Deutschland unmittelbar bildungspolitische Bedeutung gewonnen. Groß angelegte Projekte zur Entwicklung, Implementierung und Prüfung von Bildungsstandards in allen deutschsprachigen Ländern stützen sich darauf (das schweizerische Projekt Harmos (Schweizerische Konferenz der kantonalen Erziehungsdirektoren, 2007), die Arbeit des KMK-Instituts zur Qualitätsentwicklung im Bildungswesen (Köller, 2005) und Standard-Projekte in Österreich,

(Freudenthaler, Specht & Paechter, 2004)). Mit der Orientierung an anspruchsvollen Kompetenzkonzepten und -modellen soll auch der häufig anzutreffenden Kritik begegnet werden, Tests verengten den Zielhorizont von Bildungseinrichtungen und förderten oberflächliches Lernen. Ein „teaching to the test" könnte beispielsweise durchaus erwünschte Wirkungen haben, wenn ein entsprechender Test nicht nur spezifische Lerninhalte abfragt, sondern tatsächlich umfassendere Kompetenzen misst.

Die aktuell vorliegenden Standards der Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) haben den Anspruch, Erwartungen und Normen an fachliches Lernen im Kontext allgemeiner Bildungsziele zu spezifizieren. Die Leistungsstandards zeichnen sich durch verschiedene Merkmale aus:

- Es handelt sich bei diesen Standards um eine spezifische Art von curricularen Dokumenten, die Bildungsziele auf Kernbereiche fokussieren und exemplarisch operationalisieren.
- Sie konkretisieren die Ziele in Form von Kompetenzanforderungen und legen fest, über welche Kompetenzen eine Schülerin oder ein Schüler zu einem bestimmten Zeitpunkt in der Schullaufbahn verfügen muss, wenn wichtige Ziele der Schule als erreicht gelten sollen. Damit zielen sie auf kumulatives und systematisch vernetztes Lernen.
- Die Standards decken nicht die gesamte Breite des Lernbereichs bzw. Faches in allen Veränderungen ab, sondern konzentrieren sich auf einen Kernbereich. Damit legen sie fest, was für alle verbindlich ist.
- Die Kompetenzanforderungen werden in Kompetenzmodellen systematisch über verschiedene Kompetenzstufen geordnet, die Aspekte und Abstufungen von Kompetenzen darstellen.
- Konkretisiert werden sie in Aufgabenstellungen und Testverfahren, mit denen das Kompetenzniveau, das die Schülerinnen und Schüler erreicht haben, valide erfasst werden kann.

Diese Aufzählung macht deutlich, dass den Definitionen und Operationalisierungen der „Kompetenzen" von Schülerinnen und Schülern eine zentrale Bedeutung für die Ausgestaltung und Umsetzung der Bildungsstandards zukommt. Zu Merkmalen von Bildungsstandards im allgemeinen vgl. Klieme et al., 2003, S. 20ff; zur Erweiterung auf nicht fachgebundene Kompetenzen Maag Merki, 2004). Die Entwicklung von entsprechenden Kompetenzmodellen stellt eine der größten Herausforderungen für die Implementierung von Bildungsstandards in den deutschsprachigen Ländern dar.

Fragen der Ergebnismessung von Bildungsprozessen werden jedoch nicht nur im Zusammenhang mit der Praxis von Lehr-Lern-Prozessen und deren administrativer Steuerung relevant. Die Begrenzung vorhandener Messmodelle erweist sich auch als Hindernis für den Fortschritt der Bildungsforschung selbst. Die Möglichkeit, in Trainingsexperimenten, bei Unterrichtsstudien und in der Interventionsforschung Effekte festzustellen, steht und fällt mit dem Vorhandensein präziser und gültiger Messinstrumente. Insbesondere für die Prüfung differenzieller Effekte (d.h. unterschiedliche Ergebnisprofile in Abhängigkeit von Lernvoraussetzungen), für kurzfristige konzeptuelle Veränderungen oder sehr langfristige Veränderungen fehlen der Forschung häufig adä-

quate Messkonzepte, wie sich unter anderem im 2006 auslaufenden DFG-Schwerpunktprogramm „Bildungsqualität von Schule" gezeigt hat (Prenzel & Allolio-Näcke, 2006). Wie wichtig die Spezifikation der Messkonzepte in der Bildungsforschung ist, haben jüngst Gijbels, Dochy, van den Bossche und Segers (2005) am Beispiel des vor allem in der Hochschulausbildung populären „problem-based learning" gezeigt: Je nachdem, wie Problemlösekompetenz konzeptualisiert und operationalisiert wird, zeigen sich systematisch unterschiedliche Effekte. Da die empirische Bildungsforschung nicht als Disziplin konstituiert ist, sondern als ein Forschungsfeld mit Beiträgen insbesondere aus der Erziehungswissenschaft, der Psychologie und den Fachdidaktiken, müssen zur Entwicklung fundierter Kompetenzmodelle verschiedene Disziplinen und Teildisziplinen zusammenwirken. Das große Interesse der beteiligten Disziplinen wurde im April 2006 bei der Ausschreibung des Schwerpunktprogramms „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" durch die Deutsche Forschungsgemeinschaft deutlich (vgl. Klieme & Leutner, 2006), zu der mehr als achtzig Projektanträge eingingen.

1.3 Kompetenzmodelle

Bei der empirischen Erfassung von Kompetenzen stellt sich die Frage, welche Modelle die Grundlage für die Entwicklung von Messinstrumenten und für die Beschreibung von Messergebnissen bilden. Hierbei lassen sich zwei Formen von Modellen unterscheiden: *Kompetenzniveaumodelle* und *Kompetenzstrukturmodelle* (vgl. Hartig & Klieme, 2006). Beide Formen der Modellierung beziehen sich auf unterschiedliche Aspekte von Kompetenzkonstrukten, die sich gegenseitig keineswegs ausschließen sondern idealerweise ergänzen können.

Kompetenzniveaumodelle befassen sich vor allem mit der Frage, was unterschiedliche Personen können, d.h. welche spezifischen Anforderungen sie bewältigen können. Kompetenzmodelle sind vor allem bei der Ergebnismessung von Bildungsprozessen zum Zweck der Evaluation („Output") nützlich, aber auch bei der Formulierung und Untersuchung von Modellen zur Kompetenzentwicklung.

Bei Kompetenzstrukturmodellen steht hingegen eher im Mittelpunkt, wie die Bewältigung unterschiedlicher Anforderungen miteinander zusammenhängt und auf welchen und wie vielen Dimensionen interindividuelle Unterschiede in Kompetenzen angemessen beschrieben werden können. Kompetenzstrukturmodelle sind insbesondere bei Forschungsfragestellungen interessant, die sich mit der differenzierten Diagnostik von Teilkompetenzen befassen.

1.3.1 Kompetenzniveaumodelle

Bei der Messung von Kompetenzen werden typischerweise quantitative Testwerte gewonnen. Während in der psychologischen Diagnostik eine rein normorientierte Interpretation quantitativer Testwerte (z.B. IQ-Normen) dominiert, wird dies in der Bildungsforschung oft nicht mehr als ausreichend erachtet (z.B. Helmke & Hosenfeld, 2004). Neben dem Vergleich von Leistungswerten an Bezugspopulationen oder dem Vergleich von Subpopulationen (z.B. die Länder-

vergleiche in PISA) ist vielmehr von Interesse, über welche spezifischen Kompetenzen Schüler verfügen bzw. welche fachbezogenen Leistungsanforderungen sie bewältigen können. Es besteht also der Bedarf an einer kriteriumsorientierten Interpretation der quantitativen Leistungswerte (vgl. Kapitel 2; auch Klauer, 1986; Goldhammer & Hartig, in Druck). Eine inhaltliche Beschreibung der numerischen Werte auf der Kompetenzskala anhand konkreter, fachbezogener Kompetenzen ist für jeden einzelnen Punkt einer kontinuierlichen Skala in der Praxis nicht realisierbar (Beaton & Allen, 1992). Um dennoch eine kriteriumsorientierte Beschreibung der quantitativen Werte zu erzielen, wird in der Bildungsforschung oft ein pragmatischeres Vorgehen gewählt: Die kontinuierliche Skala wird in Abschnitte unterteilt, welche als Kompetenzniveaus oder Kompetenzstufen bezeichnet werden. Für diese Skalenabschnitte wird dann eine kriteriumsorientierte Beschreibung der erfassten Kompetenzen vorgenommen. Diese Definition von Abschnitten auf einer kontinuierlichen Skala ist, wenngleich oft willkürlich (Adams & Wu, 2002), ein Versuch die Beliebigkeit der Skalierung von Testwerten zu überwinden und zu nicht-beliebigen Skalen (*non-arbitrary metrics*; Embretson, 2006) zu gelangen (vgl. auch Haertel & Lorié, 2004).

Kompetenzniveaumodelle befassen sich mit der Definition von Skalenabschnitten und ihrer inhaltlichen Beschreibung. Bei den meisten Niveaumodellen wird mit Hilfe von Methoden der Item-Response-Theorie (z.B. van der Linden & Hambleton, 1997; Rost, 2004; Wilson, 2005) eine gemeinsame Skala gebildet, auf der sowohl die Kompetenzen der Personen als auch die Schwierigkeiten der Aufgaben dargestellt werden. Hierdurch können über Personen mit unterschiedlich hohen Kompetenzen Aussagen gemacht werden, welche Aufgaben sie bewältigen können und welche nicht. Auf Methoden zur Definition und Beschreibung von Kompetenzniveaus wird in Kapitel 2 dieses Bandes eingegangen.

Die kriteriumsorientierte Interpretation von Testwerten, die durch die Definition von Kompetenzniveaus möglich wird, ist insbesondere für Untersuchungen interessant, in denen der „Output" von Bildungsprozessen gemessen werden soll. Der Bezug der gemessenen Kompetenzen auf spezifische Anforderungen erlaubt Vergleiche zwischen dem empirisch beobachteten Leistungsniveau und dem als Ergebnis eines Bildungsprozesses angestrebten Niveau, das zum Beispiel Bildungsstandards formuliert wird.

1.3.2 Kompetenzstrukturmodelle

Kompetenzen werden im Sinne der hier verwendeten Arbeitsdefinition primär durch Bereiche situativer Anforderungen definiert. Um diese Anforderungen zu bewältigen, können für eine Person unter Umständen durchaus unterschiedliche Fähigkeiten nötig sein. So muss ein Individuum zur erfolgreichen Kommunikation in einer Fremdsprache über Vokabelwissen verfügen, Grammatik und Aussprache beherrschen und kulturspezifische Interaktionsregeln kennen. „Fremdsprachkompetenz" müsste sich in diesem Beispiel also nicht unbedingt als ein eindimensionales Konstrukt abbilden lassen, sondern durch verschiedene, vermutlich korrelierte Teilkompetenzen. Bei diesen Teilkompetenzen kann es sich um verschiedene Arten individueller Ressourcen handeln, die zur Bewältigung

der für ein Kompetenzkonstrukt interessierenden situativen Anforderungen notwendig oder förderlich sind, zum Beispiel spezifische Fähigkeiten und Fertigkeiten oder bereichsspezifisches Wissen.

Kompetenzstrukturmodelle befassen sich mit der Frage der Dimensionalität von Kompetenzkonstrukten. Gegenstand dieser Modelle können die Zusammenhänge zwischen Kompetenzen in verschiedenen Bereichen sein, wenn zum Beispiel eingeschätzt werden soll, für welche Bereiche es ökonomisch sinnvoll ist, separate Messwerte zu bilden. Sind zum Beispiel interindividuelle Unterschiede in verschiedenen Bereichen sehr hoch korreliert, stellt sich die Frage, ob es nicht ökonomischer ist, diese Maße bei der Datenerhebung und der Auswertung zusammenzufassen und als eine gemeinsame Skala zu behandeln. Als separate Leistungsmaße sollten dann nur solche Messungen behandelt werden, die hinsichtlich ihrer korrelativen Zusammenhänge hinreichend unabhängig voneinander sind. Bei der Entscheidung für ein bestimmtes Strukturmodell, also der Frage, wie differenziert spezifische Kompetenzen betrachtet werden sollen, muss jeweils eine Abwägung ökonomischer und theoretischer Aspekte vorgenommen werden. Kompetenzstrukturmodelle können sich jedoch auch mit der Binnenstruktur einzelner Kompetenzbereiche befassen, d.h. den zugrunde liegenden Teilkompetenzen und ihren Zusammenhängen. In diesem Fall kann ein interessierendes Kompetenzkonstrukt differenziert hinsichtlich mehrerer zugrunde liegender Teilkompetenzen – z.B. spezifische Fertigkeiten oder spezifisches Wissen – modelliert werden (z.B. Ackerman, Gierl & Walker, 2003; Reckase, 1997; McDonald, 1997).

Die Entwicklung von Kompetenzstrukturmodellen und die Analyse von Kompetenzstrukturmodellen mit mehrdimensionalen psychometrischen Modellen bietet die Chance einer – verglichen mit eindimensionalen Modellen – differenzierteren Diagnostik und zugleich einer Prüfung von Annahmen über die Struktur der erfassten Kompetenz und Teilkompetenzen (z.B. Ackerman et al., 2003; Walker & Beretvas, 2003). Kompetenzstrukturmodelle können nützlich sein, wenn das Zustandekommen und die Förderung spezifischer Kompetenzen untersucht werden soll. Für die Bildungspraxis können Erkenntnisse über eine differenzierte Struktur von Fähigkeiten, die für Personen zur Bewältigung von Anforderungen in einem spezifischen Bereich benötigt werden, die Grundlage für Lehrpläne oder Fördermaßnahmen bilden.

Die empirische Untersuchung von Kompetenzstrukturmodellen ist mit faktorenanalytischen Methoden möglich, die in der Persönlichkeitspsychologie traditionell zur Untersuchung der Struktur von Intelligenz oder von generellen Strukturen der Persönlichkeit zum Einsatz kommen. In neuerer Zeit kommen bei der mehrdimensionalen Modellierung von Kompetenzkonstrukten jedoch auch zunehmend mehrdimensionale IRT-Modelle zum Einsatz (vgl. Kapitel 2).

1.4 Beispiele für Kompetenzkonstrukte

Ein ganz wesentliches Charakteristikum des hier verwendeten Kompetenzbegriffes ist, wie weiter oben dargestellt, die Kontextabhängigkeit; in der Tradition der psychologischen Expertiseforschung wird dabei auch von *Domänen* gesprochen.

Eine Betrachtung bestehender Kompetenzmodelle zeigt, dass diese Kontexte oder auch Domänen in der schulischen Bildung sehr häufig von Schulfächern gebildet werden. So sind im Rahmen von PISA Kompetenzen erfasst und modelliert worden, die „fachbezogene Bereiche" betreffen (vgl. Leutner, Klieme, Meyer & Wirth, 2004). Ein Beispiel ist die mathematische Kompetenz, die in definiert wird als

> „[…] die Fähigkeit einer Person, die Rolle zu erkennen und zu verstehen, die Mathematik in der Welt spielt, fundierte mathematische Urteile abzugeben und Mathematik in einer Weise zu verwenden, die den Anforderungen des Lebens dieser Person als konstruktivem, engagiertem und reflektiertem Bürger entspricht. (OECD, 2003, S. 24; Übersetzung Blum et al., 2004, S. 48).

Dabei lassen sich die konkreten Anforderungen, die in den PISA-Mathematiktests realisiert werden, wie folgt charakterisieren: Anhand der Situationen, in welche die mathematischen Probleme eingebettet sind, anhand der zur Lösung notwendigen mathematischen Inhalte und anhand der Fähigkeiten, die notwendig sind, um die Probleme aus dem realen Kontext in Bezug zu mathematischen Konzepten zu setzen und damit zu einer Problemlösung zu kommen (OECD, 2003). Die Strukturierung der erfassten Leistungen erfolgt letztlich vor allem anhand der mathematischen Inhalte, welche nach vier übergreifenden Ideen („overarching ideas", OECD, 2003) unterteilt werden: *Quantität*, *Veränderung und Beziehungen*, *Raum und Form* sowie *Unsicherheit* (vgl. OECD, 2003; Blum et al., 2004).

Auch die Lesekompetenz, die neben mathematischer und naturwissenschaftlicher Kompetenz eine Hauptkomponente der PISA-Tests darstellt, kann im Sinne Weinerts (1999) als kontextspezifische, auf das Verstehen schriftlicher Texte bezogene kognitive Leistungsdisposition definiert werden. Die Anforderungen, auf die sich die PISA-Lesekompetenz bezieht, lassen sich anhand folgender Kriterien differenzieren: Die Aufgabenart, die Form und Struktur des Lesestoffs sowie der Zweck, für den der Text geschrieben wurde (OECD, 2001). Hierbei wird vor allem die Aufgabenart herangezogen, um eine Binnenstruktur der erfassten Schülerkompetenzen zu definieren. Neben einem Gesamtleistungswert werden anhand der Aufgabenart drei Subskalen gebildet, nämlich „Informationen ermitteln", „Textbezogenes Interpretieren" sowie „Reflektieren und Bewerten". Diese (Teil-)Kompetenzen werden – jedenfalls in der Sekundarstufe – nicht ausschließlich innerhalb eines einzelnen Faches gefordert und gefördert, aber die PISA-Diskussion in Wissenschaft und Politik bezieht sie doch im Wesentlichen auf den Unterricht in der Verkehrssprache. Auch in der Schulleistungsstudie DESI wird Lesekompetenz dem Fach Deutsch zugeordnet (Beck & Klieme, 2007).

Zusätzlich zu diesen eher fachbezogenen Bereichen werden im Rahmen der schulischen Bildung jedoch auch so genannte fachübergreifende Kompetenzen gemessen (Maag Merki & Grob, 2005). Hierzu gehört auch die in PISA 2003 mit erfasste Problemlösekompetenz. Diese wurde definiert als die Fähigkeit, „anwendungsbezogene fächerübergreifende Problemstellungen zu erkennen, zu verstehen und zu lösen" (Leutner et al., 2004, S. 147). Die zur Erfassung dieser Kompetenz eingesetzten Problemlöseaufgaben lassen sich systematisch nach

Problemtypen und den notwendigen Problemlöseprozessen charakterisieren. Die Problemlösetypen werden hierbei unterschieden in *Entscheidungen treffen* („decision making"), *Systeme analysieren und entwerfen* („system analysis and design") und *Fehler suchen* („trouble shooting") (vgl. Leutner et al., 2004; OECD, 2004b). Zusätzlich zu den Problemtypen lassen sich die Anforderungen der Problemlöseaufgaben hinsichtlich der zur vollständigen Lösung notwendigen Prozesse (das Problem verstehen, angemessen charakterisieren, angemessen repräsentieren, lösen, die Lösung reflektieren sowie kommunizieren) beschreiben. Diese Anforderungen stellen sich in Form von Schritten, die auf dem Weg zu einer vollständigen Aufgabenlösung bewältigt werden müssen, dar (nach OECD, 2004b). Am Beispiel der Problemlösekompetenz wird anschaulich, dass auch Konstrukte unter dem Begriff „Kompetenz" gefasst werden, für die die charakteristische Kontextabhängigkeit nicht gegeben ist oder für die zumindest eine sehr großzügige Definition von „Kontext" benötigt wird. Dies betrifft gerade solche Konstrukte, die als „fachübergreifende Kompetenzen" oder „Schlüsselkompetenzen" bezeichnet werden. Hier besteht die Gefahr, dass die Grenze zwischen „Kompetenzen" und allgemeinen kognitiven Fähigkeitskonstrukten wie Intelligenz oder auch anderen bereits existierenden Persönlichkeitskonstrukten unscharf wird (vgl. Hartig & Klieme, 2006; Klieme, 2004c).

Im Bereich der beruflichen und professionellen Expertise können Berufsfelder die entsprechenden Kontexte konstituieren, oder bestimmte Aufgabenarten und mit ihnen verbundene Anforderungen, die in verschiedenen Berufen (übergreifend) vorgefunden werden, stellen die Kontexte dar. Beispiele für Kompetenzmodelle im beruflichen Zusammenhang sind z.B. Arbeiten zum kaufmännischen Entscheidungshandeln oder zum Organisieren von Informationen in technischen Steuerungsprozessen (Breuer & Satish, 2003). Auch zu wohl definierten Aspekten der professionellen Kompetenz von Lehrpersonen (z.B. Hartinger & Fölling-Albers, 2004; Blömeke, 2003) sind Modelle entwickelt worden, etwa zur „diagnostischen Kompetenz" als Begriff dafür, inwieweit Lehrpersonen Stärken und Schwächen von Lernenden identifizieren, wie gut sie Tests und Vergleichsarbeiten selbst erstellen und auswerten und ob sie Informationen aus Leistungsvergleichen und Evaluationen angemessen interpretieren und nutzen können (Klieme & Leutner, 2006; vgl. aber auch Spinath, 2005). Ein weiteres Konstrukt im Bereich der beruflichen Expertise von Lehrpersonen ist das *pedagogical content knowledge* (Shulman, 1986), mit welchem das didaktische Fachwissen und die Fähigkeit, Fachinhalte und Unterrichtsmethoden in Beziehung zu setzen, bezeichnet wird.

2 Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle

Johannes Hartig & Nina Jude

2.1 Anlässe und Ziele der Erfassung von Kompetenzen

Der Diagnostik von Kompetenzen kommt eine Schlüsselfunktion für die Optimierung von Bildungsprozessen und für die Weiterentwicklung des Bildungswesens zu. Die in der pädagogisch-psychologischen Forschung entwickelten Verfahren zur Kompetenzmessung dienen als Grundlage für individuumsbezogene Förder-, Platzierungs- und Auswahlentscheidungen sowie für die Benotung und Zertifizierung von Lernenden. Darüber hinaus werden Ergebnisse aus empirischen Messungen auch zur Evaluation von pädagogischen Maßnahmen und Institutionen sowie für die laufende Beobachtung der Qualität von Bildungssystemen (*system-monitoring*) herangezogen. Letztlich spielt die empirische Erfassung von Kompetenzen auch in der Grundlagenforschung immer dann eine entscheidende Rolle, wenn – wie etwa in der Unterrichtsforschung – Bedingungen und Fördermöglichkeiten der Kompetenzentwicklung untersucht werden.

Trotz der großen Rolle von Kompetenzmessungen wird in Bildungspraxis und Bildungspolitik häufig noch unterschätzt, wie anspruchsvoll die empirische Erfassung von Kompetenzen aus theoretischer und methodischer Perspektive ist. Die Entwicklung sowohl theoretisch als auch empirisch fundierter Kompetenzmodelle als Ausgangspunkt für die Entwicklung adäquater Messverfahren stellt immer noch eine Herausforderung dar (z.B. Hartig & Klieme, 2006).

Nach einigen kurzen technischen Begriffsdefinitionen behandelt dieses Kapitel zunächst allgemeine Qualitätsanforderungen, die an Messverfahren zur Erfassung von Kompetenzen zu stellen sind. Anschließend wird auf verschiedene kritische Fragen im Zusammenhang mit der Entwicklung spezifischer Messinstrumente eingegangen, z.B. mögliche Strategien der Generierung von Aufgabeninhalten und verschiedene mögliche Antwortformate. Im letzten Abschnitt werden kurz die Funktion und mögliche Spezifizierungen von psychometrischen Modellen behandelt, mit denen Testdaten ausgewertet und auf die Kompetenzen der getesteten Personen geschlossen werden sollen.

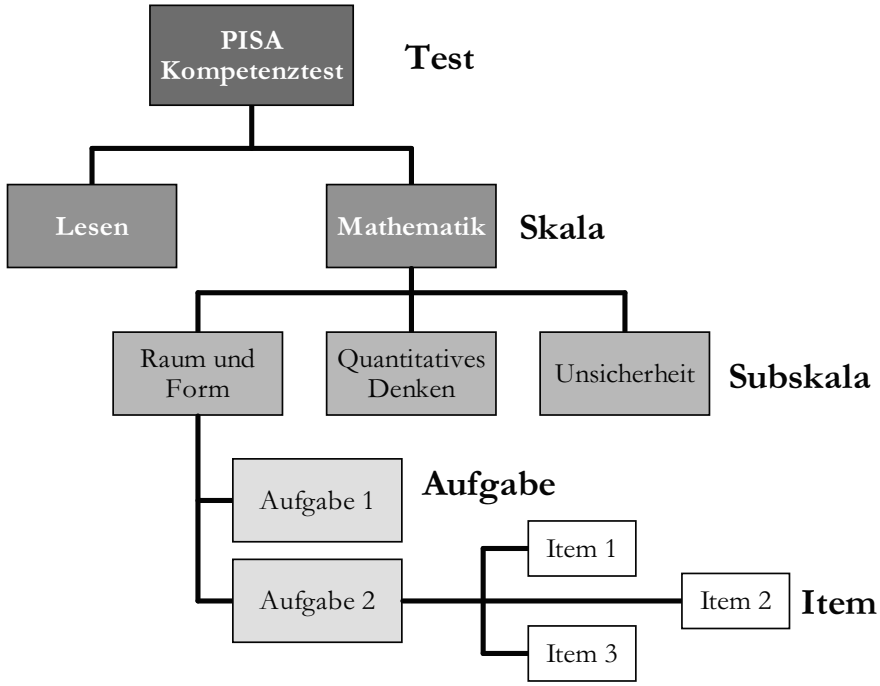
In den folgenden Abschnitten werden eine Reihe technischer Begriffe wiederkehrend gebraucht, die in verschiedenen Kontexten teilweise unterschiedliche Bedeutungen haben können. Die Verwendung dieser Begriffe im vorliegenden Text soll hier einleitend kurz erläutert werden.

Assessment

Im Englischen bedeutet *assessment* allgemein eine Einschätzung, Beurteilung oder Bewertung und bezieht sich nicht nur auf das Beurteilen von Personen, sondern auch von Situationen, Organisationen usw. Der Begriff ist im Kontext

pädagogischer und psychologischer Diagnostik in unterschiedlichen Fachzusammenhängen ins Deutsche übernommen worden (z.B. *assessment center*, *large scale assessment*), wo er eine entsprechend breite und ungenaue Bedeutung hat. Im vorliegenden Band wird *Assessment* als übergeordneter Begriff für das Beurteilen von Personen hinsichtlich bestimmter Merkmale oder bestimmter Eignungen im gleichen allgemeinen Sinn wie „Diagnostik" gebraucht. Assessment kann standardisierte Testverfahren (s.u.) beinhalten, aber auch weniger stark standardisierte Beurteilungsmethoden wie Interviews, Portfolios oder biographische Befragungen fallen unter den Begriff Assessment.

Abbildung 2.1: Schematische Darstellung einer Hierarchie von Test, Skala, Aufgabe und Item am Beispiel von Testinhalten in PISA 2003 (vgl. z.B. Prenzel et al., 2004).



Test, Fragebogen, Skala und Subskala

Unter einem *Test* wird im Rahmen der pädagogisch-psychologischen Diagnostik „ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung" verstanden (Lienert, 1969, S. 7). Diese Definition schließt sowohl Leistungstests, in denen vorgegebene Probleme gelöst werden müssen, als auch Fragebögen, in denen eine subjektive Einschätzung der eigenen Person verlangt wird, ein. Im vorliegenden Kontext werden mit *Test* lediglich Leistungstests bezeichnet, Verfahren zur Selbsteinschätzung im Unterschied dazu mit *Fragebögen*.

Die Begriffe *Skala, Subskala*, *Aufgabe* und *Item* beziehen sich auf Teile eines Tests und stehen in einer hierarchischen Beziehung zueinander. Diese Hierar-

chie wird in Abbildung 2.1 am Beispiel von Inhalten des in PISA 2003 eingesetzten Tests veranschaulicht.

Ein Test kann also mehrere Merkmale oder Konstrukte erfassen, d.h. aus der Auswertung können mehrere Messwerte resultieren. Im Unterschied dazu erfasst eine *Skala* nur ein Merkmal, das Ergebnis einer Skala kann in der Regel durch einen einzelnen Messwert ausgedrückt werden. Eine Skala kann aus *Subskalen* bestehen, die differenziertere Facetten des in der Skala erfassten Merkmals messen. Subskalen können sowohl separat ausgewertet werden als auch zu der übergeordneten Skala zusammengefasst werden.

Aufgabe und Item

In einem Leistungstest bezeichnet eine *Aufgabe* eine inhaltlich zusammenhängende, nicht aufteilbare Einheit, in der ein Problem oder eine bestimmte Einheit von Material (z.B. ein Lesetext) vorgegeben wird. Ein *Item* hingegen stellt die kleinste Analyseeinheit eines Tests dar, die bei der Auswertung berücksichtigt wird. Typischerweise ist ein Item dabei gleichbedeutend mit einer einzelnen Frage oder Teilfrage. Eine Testaufgabe kann also aus mehreren Items bestehen. Dies muss aber nicht der Fall sein. Wenn zu einer Aufgabe nur eine einzelne Antwort zu geben ist, sind Aufgabe und Item identisch. In Fragebögen, in denen per Definition keine „Aufgaben" enthalten sind, ist in der Regel eine Menge von Items direkt einer Skala untergeordnet.

2.2 Anforderungen an empirische Messverfahren

An ein nach wissenschaftlichen Kriterien konstruiertes diagnostisches Verfahren werden eine Reihe von Qualitätsansprüchen gestellt. Nur wenn diese *Gütekriterien* erfüllt sind, können mit einem Messverfahren gewonnene Ergebnisse zur empirischen Untersuchung wissenschaftlicher Fragestellungen herangezogen werden, und nur dann dürfen auf Basis der Ergebnisse Konsequenzen für Individuen oder Institutionen gezogen werden. An Tests zur Erfassung von Kompetenzen sind zunächst dieselben gängigen Gütekriterien anzulegen, die in der pädagogisch-psychologischen Diagnostik generell Verwendung finden: *Objektivität*, *Reliabilität* und *Validität*. Diese drei Konzepte und ihre Bedeutung für die Diagnostik von Kompetenzen werden in den folgenden Abschnitten erläutert. Zusätzlich wird auf eine speziell für die Kompetenzdiagnostik relevante Anforderung eingegangen, nämlich die Notwendigkeit einer *kriteriumsorientierten Testwertinterpretation*.

2.2.1 Objektivität

Mit *Objektivität* eines Tests ist vor allem gemeint, dass das Testergebnis einer Person nur von den Merkmalen der Person abhängt, und nicht durch den Testleiter oder die Testsituation beeinflusst wird. Häufig wird in die Definition des Gütekriteriums Objektivität auch noch einbezogen, dass ein Testergebnis und die Interpretation desselben nicht von der Person abhängen, die die Auswertung und Interpretation vornimmt. Entsprechend der verschiedenen Kriterien

zur Beurteilung der Objektivität eines Tests kann differenziert werden in Durchführungs-, Auswertungs- und Interpretationsobjektivität eines Tests.

Die geläufigste Strategie, die Objektivität eines diagnostischen Verfahrens sicherzustellen, ist die gründliche Standardisierung und Dokumentation aller Schritte, die bei Testdurchführung, Auswertung und Interpretation nötig sind. Ist eine derartige Standardisierung gegeben und wird die Testung durch einen qualifizierten Testleiter durchgeführt, wird Objektivität zumeist als gegeben betrachtet. Wenngleich Objektivität als notwendige Voraussetzung für Reliabilität und Validität betrachtet werden kann (Rost, 2004), wird dieses Gütekriterium in der psychologischen Diagnostik weit weniger häufig thematisiert und beforscht.

Für standardisierte psychologische Tests, die zudem meistens nur geschlossenen Antwortformate enthalten (s.u., Abschnitt 2.3.3 „Antwortformate"), mag Objektivität in der Tat kein kritisches Thema sein. In pädagogischen Kontexten werden jedoch auch eine Reihe anderer Datenquellen herangezogen, von denen auf interindividuelle Unterschiede in Kompetenzen geschlossen werden soll, z.B. Verhaltensbeobachtungen, Beurteilungen von frei formulierten Texten oder Portfolios. Angesichts der Definition von Kompetenzen als kontextabhängige, „realitätsnahe" Konstrukte erscheint eine solche größere Breite von Methoden zur empirischen Erfassung angemessen. Da die Auswertung von offenen Antwortformaten (s.u.) in der Regel jedoch das Einschätzen der beobachteten Verhaltensweisen durch Beurteiler (engl. *rater*) erfordert, ist die Frage nach der Objektivität der Messungen hier durchaus von Bedeutung. Wenn ein diagnostischer Prozess Bewertungen durch Beurteiler einbezieht, ist es unerlässlich, eine gründliche Dokumentation und Anleitung für den Beurteilungsprozess zu erstellen und die Beurteiler auf dieser Basis zu schulen. Ebenso unerlässlich ist es, wenigstens stichprobenweise die Übereinstimmung zwischen mehreren Beurteilern, die dieselben Personen beurteilen, zu untersuchen. Eine mangelnde Beurteilerübereinstimmung stellt einen Hinweis darauf dar, dass die Beurteilungskriterien, der Beurteilungsprozess oder dessen Dokumentation einer Revision bedürfen.

Es ist im Kontext des vorliegenden Bandes anzumerken, dass der Einsatz von Computertechnologie auch im Zusammenhang mit Beurteilerprozessen durch menschliche Beurteiler – die im engeren Sinne nicht Bestandteil eines „Technologiebasierten Assessment" sind – deutliche Vorteile bietet. So können Urteile in der Phase des Beurteilertrainings in einer netzwerkbasierten Umgebung wechselseitig eingesehen werden, und auch in späteren Phasen einer Untersuchung ermöglicht eine zentrale Sammlung der Urteile eine schnelle Identifikation von beurteiler- oder aufgabenspezifischen Übereinstimmungsproblemen.

Im Zusammenhang mit dem Gütekriterium der Objektivität soll hier kurz der Begriff des „objektiven Tests"[1] im Kontrast zu selbst eingeschätzten Kompetenzen erwähnt werden. In diesem Fall bezieht sich „objektiv" nicht auf Ob-

[1] Mit objektiven Tests sind hier lediglich Tests gemeint, in denen *objektivierbare* – typischerweise leistungsbezogene – Verhaltensweisen erfasst werden. Nicht gemeint sind objektive Tests im Sinne von Cattell (z.B. Cattell & Warburton, 1967), die dadurch gekennzeichnet sind, dass das erfasste Konstrukt der gestesteten Person nicht erkenntlich sein soll.

jektivität im engeren Sinne des hier beschriebenen Gütekriteriums, sondern auf eine Objektivierbarkeit von erfassten Testleistungen im Unterschied zu „subjektiven" Selbsteinschätzungen. Auf die Verwendung von Selbsteinschätzungen wird später in diesem Kapitel eingegangen.

2.2.2 Reliabilität

Reliabilität bezeichnet die Messgenauigkeit eines Tests. Dieses Gütekriterium spielt vor allem im Kontext der so genannten Klassischen Testtheorie (KTT) eine zentrale Rolle. In der KTT existieren verschiedene statistische Verfahren zur empirischen Einschätzung der Reliabilität. Die gemeinsame Grundidee dieser Methoden ist, dass einem einzelnen Messwert ein „wahrer Wert" im interessierenden Merkmal zugrunde liegt; jede Messung ist jedoch mit einem als „zufällig" behandelten Messfehler behaftet. Die Methoden der KTT dienen dazu einzuschätzen, welcher Anteil der beobachteten Streuung der mit einem Test ermittelten Messwerte tatsächlich auf Unterschiede im interessierenden Merkmal zurückzuführen ist. Dieser Anteil „wahrer" Merkmalsvarianz an der beobachteten Testwertvarianz ist per Definition die Reliabilität eines Tests. Ein häufiger Weg zur Reliabilitätsschätzung ist die Bestimmung der so genannten *internen Konsistenz*. Hierbei werden hohe korrelative Zusammenhänge zwischen Aufgaben oder Teiltests als Hinweis auf eine hohe Messgenauigkeit interpretiert.

Die Reliabilitätsbestimmung in der KTT, insbesondere über die interne Konsistenz, geht davon aus, dass dem jeweiligen Test ein einziges quantitatives Merkmal zugrunde liegt. Diese Annahme ist für Kompetenztests, deren Konstrukte sich aus einem spezifischen Anforderungskontext definieren, keineswegs selbstverständlich. Es ist daher jeweils gründlich zu prüfen, ob Koeffizienten wie Cronbachs Alpha oder die Halbtest-Reliabilität[2] (s. z.B. Amelang & Schmidt-Atzert, 2006) geeignete Maße zur Bestimmung der Messgenauigkeit eines Kompetenztests sind. Diese Frage gilt auch für Reliabilitätsmaße im Rahmen der Item-Response-Theorie (IRT)[3], soweit sie aus Modellen mit einem eindimensionalen Merkmal abgeleitet sind. Falls einem Kompetenzkonstrukt eine mehrdimensionale Struktur zugrunde liegt und ein entsprechendes psychometrisches Modell formuliert ist, können Schätzungen der Reliabilität alternativ darüber erfolgen, wie gut dieses Modell Variation in den Lösungen der einzelnen Aufgaben erklärt (s.u., Abschnitt 2.4 „Psychometrische Modelle zur Messung der Kompetenz").

Eine ebenfalls geläufige Methode zur Einschätzung der Reliabilität, für die die Annahme eines eindimensionalen Konstrukts nicht notwendig ist, ist die

---

[2] Schätzungen der Reliabilität in der Klassischen Testtheorie basieren im Wesentlichen auf den Zusammenhängen zwischen Teilen des Tests. So ist Cronbach's Alpha abhängig von der durchschnittlichen Interkorrelation aller Items, die Halbtest-Reliabilität basiert auf der Korrelation zwischen zwei Testhälften (z.B. Schermelleh-Engel & Werner, in Druck).

[3] In Modellen der IRT werden Antwortwahrscheinlichkeiten für die Items eines Tests als Funktion der zu messenden Merkmale modelliert. Schätzungen der Reliabilität hängen auch hier von den Zusammenhängen der Items ab, die ein gemeinsames Merkmal messen sollen.

mehrfache Messung mit demselben Test. Der Zusammenhang zwischen den beiden Messungen wird dann als Maß für die Reliabilität herangezogen (*Retest-Reliabilität*). Wenn eine solche wiederholte Messung in einem Zeitraum stattfindet, über den die interessierende Kompetenz als stabil betrachtet wird, kann die Retest-Reliabilität eine angemessene Methode zur Schätzung der Messgenauigkeit darstellen.

2.2.3 Validität

Validität („Gültigkeit") bezeichnet den Umfang, in dem ein Test tatsächlich dasjenige Merkmal erfasst, das er erfassen soll. Damit stellt die Validität das wichtigste Gütekriterium zur Beurteilung eines diagnostischen Verfahrens dar. In der häufig zitierten Begriffsbestimmung von Messick (1989) wird Validität definiert als „an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment." (S. 13). Validität stellt also die notwendige Legitimation für das Ziehen individueller und institutioneller Konsequenzen aus einer Testanwendung dar. In der Literatur haben sich mittlerweile eine Reihe verschiedener Konzepte und Methoden zur Prüfung der Validität etabliert. Diese unterschiedlichen Methoden haben zu einer Vielfalt „unterschiedlicher Validitäten" geführt, die sich auf sehr unterschiedliche Kriterien beziehen (vgl. Borsboom, Mellenbergh & van Heerden, 2004). Hier soll lediglich auf die Bedeutung der Begriffe *Konstruktvalidität*, *Kriteriumsvalidität* und *Inhaltsvalidität* für die Messung von Kompetenzen eingegangen werden.

Konstruktvalidität

Bei der Untersuchung der Konstruktvalidität (Cronbach & Meehl, 1955) soll die Frage, was der Test misst, auf Basis theoretischer Annahmen über das zu messende Konstrukt beantwortet werden. Aus den Annahmen über das Konstrukt werden Vorhersagen abgeleitet, wie Messwerte des Konstrukts mit anderen Variablen zusammenhängen sollten. Die Gesamtheit der Aussagen über Beziehungen zwischen verschiedenen Konstrukten konstituiert ein so genanntes *nomologisches Netzwerk*, in dem das zu messende Konstrukt verortet werden kann. Dieses nomologische Netzwerk besteht aus der Gesamtheit der auf theoretischer Ebene angenommenen Zusammenhänge zwischen dem interessierenden Konstrukt und anderen theoretischen Konstrukten. Die Untersuchung der Konstruktvalidität erfolgt durch die empirische Prüfung der aus dem nomologischen Netzwerk abgeleiteten Vorhersagen, wozu Daten mit dem interessierenden Messinstrument und anderen relevanten Variablen erhoben werden (vgl. Hartig, Frey & Jude, in Druck).

Das Konzept der Konstruktvalidität trifft den Kern der Frage „was ein Test misst", wenn ein Test ein theoretisches, von der Art der Messung unabhängiges Konstrukt erfassen soll („whenever a test is to be interpreted as a measure of some attribute or quality which is not ‚operationally defined'"; Cronbach & Meehl, 1955, S. 282). Für Kompetenzkonstrukte ist dieser Ansatz nur bedingt geeignet. Die primäre Intention der Kompetenzmessung ist die Vorhersage erfolgreichen Handelns in einem interessierenden Bereich von Situationen. Ob

es möglich und erstrebenswert ist, für ein einzelnes kontextabhängig definiertes Konstrukt ein elaboriertes nomologisches Netzwerk zu konstruieren, muss im Einzelfall entschieden werden. Während dies bei gut beforschten und homogenen Konstrukten wie Lesekompetenz noch vorstellbar erscheint, ist dies z.B. bei stark berufsbezogenen, breiteren Kompetenzkonstrukten weniger viel versprechend.

Kriteriumsvalidität

Im Mittelpunkt der Kriteriumsvalidität steht die pragmatische Frage, inwieweit mit einem Test Verhalten außerhalb der Testsituation vorhergesagt werden kann. Es geht hierbei also weniger darum, *was* ein Test misst, sondern wie gut er sich praktisch bewährt. Diese Fragestellung ist für Tests zur Erfassung von Kompetenzen fraglos zentral, da mit dem Konstrukt der Kompetenz ja gerade das Versprechen einer höheren Realitätsnähe und einer besseren Prognose verbunden ist (vgl. Kapitel 1). Es erscheint daher angemessen, bei der Entwicklung von Kompetenztests auch frühzeitig Untersuchungsdesigns und Kriterien zu entwickeln, mit denen die Vorhersagegüte geprüft werden kann.

Inhaltsvalidität

Das Kriterium der Inhaltsvalidität wird in der psychologischen Diagnostik gegenüber der Konstrukt- und Kriteriumsvalidität eher vernachlässigt. Es bezieht sich darauf, ob die Test- und Aufgabeninhalte den interessierenden Merkmals- oder Verhaltensbereich, der das zu messende Konstrukt definiert, gut repräsentieren. Für Kompetenztests erscheint diese Frage ausgesprochen zentral, da Kompetenzkonstrukte primär durch einen bestimmten Situations- und Merkmalsbereich definiert werden.

Die Prüfung der Inhaltsvalidität eines Tests erfolgt typischerweise durch Urteile von Experten für die jeweilige Inhaltsdomäne. Gerade im Bildungswesen ist die Bedeutung der Inhaltsvalidität eines Tests offensichtlich. Wenn etwa untersucht werden soll, inwieweit die im Curriculum für ein bestimmtes Fach definierten Lehrziele erreicht wurden, muss ein Test diese Ziele hinreichend gut abbilden. Ob dies der Fall ist, kann nicht durch die Untersuchung von Schülern und durch empirische Zusammenhänge mit anderen Variablen beantwortet werden, sondern nur mit Hilfe von Expertenurteilen, z.B. von Fachdidaktikern oder von für das Curriculum verantwortlichen Akteuren im Bildungsprozess.

Zusammenfassung

Eine Konstruktvalidierung im Sinne der Prüfung eines nomologischen Netzwerkes erscheint für Kompetenzkonstrukte weniger angemessen als für primär theoretisch definierte Konstrukte. Hingegen stellen sowohl die Untersuchung der Kriteriums- als auch der Inhaltsvalidität wichtige Fragestellungen bei der Entwicklung von Kompetenztests dar. Beide befassen sich – aus unterschiedlichen Perspektiven – mit der Frage, inwieweit die Ergebnisse aus einem Test auf die Realität außerhalb der Testsituation generalisierbar sind. Die Frage der Kriteriumsvalidität betrifft die praktische Bewährung eines Messinstrumentes bei der Vorhersage erfolgreichen Handelns. Die Frage der Inhaltsvalidität betrifft die Definition eines Kompetenzkonstrukts. Hier wird untersucht, ob der inte-

ressierende Situations- und Verhaltensbereich durch einen Test hinreichend gut abgebildet wird. Abschließend sei die wechselseitige Ergänzung von Kriteriums- und Inhaltsvalidität anhand eines Zitats von Cronbach & Meehl illustriert:

> "In predictive or concurrent validity [*Kriteriumsvalidität*], the criterion behavior is of concern to the tester, and he may have no concern whatsoever with the type of behavior exhibited in the test. (…) Content validity is studied when the tester *is* concerned with the type of behavior involved in the test performance." (Cronbach & Meehl, 1955, S. 282, Hervorhebung im Original)

2.2.4 Kriteriumsorientierte Testwertinterpretation

Die vorangehend genannten Anforderungen an Objektivität, Reliabilität und Validität werden an jeden psychologischen oder pädagogischen Test oder Fragebogen gestellt. Bei der Erfassung von Kompetenzen, von der sich ökologisch validere Informationen erhofft werden, kommt eine spezifische Anforderung hinzu: die Möglichkeit, auf Basis der Messwerte *kriteriumsbezogene* Aussagen machen zu können.

Die meisten psychologischen und pädagogischen Tests werden unter Zuhilfenahme *sozialer Bezugsnormen* interpretiert: Eine individuell gemessene Merkmalsausprägung wird in Relation zur Merkmalsverteilung in einer Vergleichsgruppe (z.B. Schüler derselben Jahrgangsstufe) als hoch oder niedrig eingeschätzt – d.h. als z.B. „unterdurchschnittlich" oder „überdurchschnittlich". Bekanntestes Beispiel für eine derartige bezugsnormorientierte Testwertinterpretation sind Skalen für viele Intelligenztests, die so normiert werden, dass 100 dem Durchschnitt und 15 Punkte einer Standardabweichung entsprechen. Mit *kriteriumsorientierter Testwertinterpretation* ist im Unterschied zu einer Bezugsnormorientierung gemeint, dass ein Testergebnis dahingehend interpretiert wird, ob ein bestimmtes, vorher definiertes *Kriterium* erreicht wird (vgl. Klauer, 1987). Ein derartiges Kriterium kann in der pädagogischen Diagnostik z.B. durch Bildungsstandards definiert werden. Die Einschätzung der Erreichung eines Kriteriums erfolgt unabhängig von Bezugsnormen, d.h. unabhängig davon, wie viele der anderen untersuchten Personen das Kriterium erreichen.

Eine kriteriumsorientierte Testwertinterpretation ist eine Voraussetzung für die Beschreibung und Bestimmung von *Kompetenzniveaus* (vgl. Kapitel 1). Bei der Definition von Kompetenzniveaus wird beschrieben, welche konkreten Anforderungen eine getestete Person mit einer bestimmten Ausprägung einer Kompetenz wahrscheinlich bewältigen kann, und welche (noch) nicht. Dieses Vorgehen ist für Kompetenzen im Unterschied zu anderen Fähigkeits- oder Leistungskonstrukten besonders wichtig, da der enge Bezug zu konkreten Situationen und Handlungskontexten ein definierendes Merkmal von Kompetenzen darstellt. Ohne eine kriteriumsorientierte Testwertinterpretation können z.B. keine substanziellen Aussagen über gesamte Populationen gemacht werden, wie sie für ein System-Monitoring notwendig sind (z.B., ob ein Bildungsstandard in einer bestimmten Jahrgangsstufe in einem Bundesland erreicht wurde).

Zur Definition und Beschreibung von Kompetenzniveaus werden häufig Messmodelle der Item-Response-Theorie (IRT) herangezogen, da diese eine Beschreibung von Aufgabenschwierigkeiten und Personenmerkmalen auf derselben

Skala erlauben (z.B. Rost, 2005). Die Beschreibung von Kompetenzniveaus kann im einfachsten Fall durch eine Post-Hoc-Analyse der empirisch ermittelten Aufgabenschwierigkeiten, oder unter Einbezug von systematisch beschriebenen Aufgabenanforderungen erfolgen (s.u. Abschnitt 2.3.4, vgl. auch Hartig, 2007; Hartig & Klieme, 2006). Auf die psychometrischen Methoden zur Unterstützung bei der Definition von Kompetenzniveaus wird später in diesem Kapitel ausführlicher eingegangen.

2.3 Konstruktion von Messinstrumenten

Im Kontext des vorliegenden Bandes kann nicht im Detail auf Prinzipien und Techniken der Test- und Aufgabenkonstruktion eingegangen werden, zumal sich diese auch je nach Inhaltsbereich deutlich unterscheiden können. Bei der Entwicklung eines spezifischen Messverfahrens auf Basis der Definition eines zu erfassenden Konstruktes oder einer spezifischen diagnostischen Forschungsfragestellung sind jedoch zunächst eine Reihe grundlegender Fragen zu beantworten, mit denen sich die folgenden Abschnitte befassen werden. Zunächst wird die Frage der Verwendung von Selbsteinschätzungen zur Erfassung von Kompetenzen behandelt. Anschließend werden verschiedene mögliche allgemeine Strategien zur Generierung und Auswahl von Aufgabeninhalten skizziert. In zwei weiteren Abschnitten wird kurz auf unterschiedliche Antwortformate und auf die Definition von schwierigkeitsbestimmenden Aufgabenmerkmalen eingegangen.

2.3.1 Verwendung von Selbstbeschreibungen

Ein bei der empirischen Erfassung von Kompetenzen häufig kontroverser Punkt ist die Verwendung von durch die untersuchten Personen selbst eingeschätzten Kompetenzen. Derartige Befragungen erfreuen sich großer Beliebtheit, da sie verglichen mit objektiven Messverfahren ungleich ökonomischer sind. In bestimmten Untersuchungs-Settings, in denen der Einsatz von Tests aus ökonomischen oder organisatorischen Gründen nicht möglich ist, sehen Forschende in Selbstbeschreibungen den einzigen Weg, sich den interessierenden Kompetenzen empirisch zu nähern. So werden etwa universitäre Lehrveranstaltungen oder andere Bildungsmaßnahmen evaluiert, indem die Teilnehmer um eine Einschätzung der in dieser Maßnahme erworbenen Kompetenz gebeten werden (z.B. Flexer & Baer, 2005).

Ungeachtet nachvollziehbarer pragmatischer Gründe für den Einsatz von Selbsteinschätzungen muss hinterfragt werden, ob mit derartigen Einschätzungen tatsächlich dasselbe gemessen wird wie mit objektiven Tests. Nicht umsonst werden Selbsteinschätzungen häufig als Maße für „wahrgenommene Kompetenz“ (*perceived competence*) oder „Selbstkonzept“ (*self-concept*) bezeichnet und nicht als „Kompetenz“. Auch in der deutschsprachigen schulischen Bildungsforschung gibt es eine deutliche Differenzierung zwischen *Fähigkeiten* auf der einen und fähigkeitsbezogenen oder „akademischen“ *Selbstkonzepten* auf der anderen Seite. Während erstere typischerweise über standardisierte Tests erfasst werden, werden letztere durch Selbsteinschätzungen in Fragebögen erhoben. Das Selbstkonzept stellt ein eigenes, extensiv beschriebenes und beforschtes

Konstrukt dar (z.B. Marsh, Trautwein, Lüdtke, Köller & Baumert, 2005; 2006). Aus der Forschung zu fähigkeitsbezogenen Selbstkonzepten, aber auch aus Untersuchungen zu Selbsteinschätzungen allgemeiner Intelligenz und anderer Fähigkeitskonstrukte ist bekannt, dass Testergebnisse und Selbsteinschätzungen positiv, aber allenfalls moderat miteinander zusammen hängen (z.B. Hacker, Bol, Horgan & Rakow, 2000; Kruger & Dunning, 1999; Tousignant & Des-Marchais, 2002).

Sowohl aufgrund empirischer Befunde als auch angesichts einer separaten theoretischen Konzeptualisierung fähigkeitsbezogener Selbstkonzepte erscheint es nicht zu rechtfertigen, Selbsteinschätzungen zu einem Kompetenzkonstrukt als unmittelbaren Indikator dieser Kompetenz zu behandeln. Dies bedeutet nicht, dass die in Selbsteinschätzungen enthaltenen Informationen generell uninteressant oder unnütz wären; es muss lediglich beachtet werden, dass es sich hierbei nicht um eine Messung der interessierenden Kompetenz selbst handelt, sondern um ein diesbezügliches Selbstkonzept. In Forschung zu letzterem sind fragebogenbasierte Selbsteinschätzungen naturgemäß die Datenquelle der Wahl. In Fällen, in denen der Einsatz von Tests nicht realisierbar ist, kann die Erhebung von Selbsteinschätzungen zumindest Hinweise auf die interessierenden Kompetenzen liefern, da in der Regel ein bedeutsamer positiver, wenngleich nicht sehr enger Zusammenhang zwischen Kompetenz und dem zugehörigen Selbstkonzept angenommen werden kann. In diesen Fällen wird jedoch keine direkte *Kompetenzmessung* vorgenommen, sondern über eine korrelierte Variable, nämlich das Selbstkonzept, indirekt auf die Kompetenz geschlossen.

Ein weiteres grundlegendes Problem bei der Verwendung von Selbstberichten besteht auch darin, dass kaum ein Bezug zu konkreten Anforderungen möglich ist, die eine befragte Person bewältigen kann – es ist also keine kriteriumsorientierte Interpretation der erfassten Werte möglich. Dies liegt unter anderem daran, dass Selbstberichte durch die soziale Bezugsnorm beeinflusst werden, die die befragten Personen ihren Antworten implizit zugrunde legen. Wenn zum Beispiel eine Selbsteinschätzung mit der Frage „Wie gut können Sie die Tätigkeit XY ausführen“ erhoben wird, werden die meisten Antwortenden einen Vergleich zu Personen in ihrem eigenen Umfeld anstellen und damit implizit die Frage beantworten „wie gut kann ich die Tätigkeit im Vergleich zu anderen ausführen?“. Befragte, die objektiv die gleichen Anforderungen bewältigen können, können so zu unterschiedlichen Antworten gelangen, wenn sich ihre soziale Bezugsnorm unterscheidet.

Im Kontext des vorliegenden Bandes werden Selbsteinschätzungen ausdrücklich nicht als Messinstrumente zur Erfassung von Kompetenzen betrachtet. Mit „Messinstrumenten“ oder „Tests“ sind Verfahren gemeint, die auf objektivierbaren Verhaltensdaten basieren.

2.3.2 Strategien der Testkonstruktion: Generierung und Auswahl von Test- und Aufgabeninhalten

Die zentrale erste Frage bei der Konstruktion eines neuen diagnostischen Instrumentes ist, wie die Test- und Aufgabeninhalte definiert, eingegrenzt und ausgewählt werden. Die Auswahl der Testinhalte definiert letztlich die Natur des gemessenen Konstrukts. Im Folgenden soll auf vier generelle Strategien zur

Ableitung von Testinhalten eingegangen werden: Externale Konstruktion, deduktive Konstruktion, induktive Konstruktion und Kriteriumssampling. In der Praxis kommen häufig auch Mischformen dieser Strategien zum Einsatz.

Externale Konstruktion

Bei der externalen Konstruktion steht das Ziel im Vordergrund, ein bestimmtes externes Kriterium, z.B. die Zugehörigkeit zu einer bestimmten Gruppe, vorherzusagen. Die Herkunft der Iteminhalte ist dabei sekundär, idealtypisch ist eine möglichst breite und heterogene Menge von Items (sowohl Testaufgaben als auch Fragebogenitems). Diese Items werden dann empirisch darauf hin untersucht, ob sich die interessierenden Gruppen darin unterscheiden. Die Items, hinsichtlich derer sich die Gruppen am deutlichsten unterscheiden, werden zu einem „Messinstrument" zusammengestellt. Mit diesem Instrument soll dann bei zukünftig zu untersuchenden Personen die Gruppenzugehörigkeit vorhergesagt werden. Beispiele für ein derartiges Vorgehen sind Verfahren zur Klassifikation von Patienten zu bestimmten psychiatrischen Diagnosegruppen (Hathaway & McKinley, 1951) oder zur Trennung erfolgreicher vs. erfolgloser Bewerber auf Studien- oder Arbeitsplätze.

Bei einer externalen Testkonstruktion ist allein die pragmatische Frage der optimalen Vorhersage eines Kriteriums von Interesse, theoretische Annahmen über gemessene Konstrukte spielen keine Rolle. Im Rahmen der Kompetenzdiagnostik kommt dieses Vorgehen nicht in Betracht.

Deduktive Konstruktion

Die *deduktive* oder *rationale* Testkonstruktion geht im Gegensatz zur externalen Konstruktion primär von einem theoretischen Konstrukt aus; dieses Vorgehen stellt ein Ideal für die psychologische Diagnostik dar. Es werden theoretische Annahmen darüber gemacht, wie sich Personen in bestimmten Bereichen beschreiben und unterscheiden lassen – z.B. im Hinblick auf psychische Prozesse oder neurophysiologische Strukturen. Auf Basis der Beschreibung des theoretischen Konstrukts werden mögliche beobachtbare Indikatoren abgeleitet, in denen sich interindividuelle Unterschiede im Konstrukt möglichst stark niederschlagen sollten. Diese theoriebasierte Auswahl der Indikatoren kann z.B. durch Expertenurteile abgesichert werden. Die ausgewählten Indikatoren bilden dann ein Messinstrument für das interessierende Konstrukt, das typischerweise in anschließenden empirischen Untersuchungen auf seine Validität geprüft wird. Beispiele für rational basierte Verfahren sind in der Persönlichkeitspsychologie sowohl im Leistungsbereich als auch für nicht leistungsbezogene Persönlichkeitskonstrukte zahlreich (vgl. Amelang & Schmidt-Atzert, 2006).

Eine deduktive Testkonstruktion stellt einen anspruchsvollen und viel versprechenden Ansatz dar, wenn eine hinreichende theoretische Fundierung existiert. Für Kompetenzkonstrukte ist ein deduktives Vorgehen vorstellbar, wobei zu berücksichtigen ist, dass die Messung einer Kompetenz der Erfassung eines mehrdimensionalen Konstrukts auf Personenseite entsprechen kann (vgl. Kapitel 1).

Induktive Konstruktion

Die induktive Test- oder Skalenkonstruktion ist, wie die externale Konstruktion, ein weitgehend theoriefreies Vorgehen. Sie ist eng mit der statistischen Analysemethode der Faktorenanalyse verbunden. Die zugrunde liegende Idee ist, dass Items, die miteinander korrelieren, ein gemeinsames Konstrukt erfassen. Um diese Konstrukte zu identifizieren und – im selben Schritt – die Items zu ihrer Messung zusammenzustellen, wird mit faktorenanalytischen Methoden die Struktur einer Menge von Items untersucht. Nach der Bestimmung der Anzahl relevanter Faktoren (zu Methoden z.B. Moosbrugger & Hartig, 2002) werden diejenigen Items, die gemeinsam auf einem Faktor laden, zu einer Skala zusammengestellt. Die Natur des durch diese Skala erfassten Konstrukts ergibt sich post hoc aus den Inhalten der Items. Prominente Beispiele für ein induktives, faktorenanalytisches Vorgehen im Bereich der psychologischen Diagnostik sind im Leistungsbereich die „Primary Mental Abilities" von Thurstone (1941) und im nicht leistungsbezogenen Bereich Fünf-Faktoren-Modell mit den Persönlichkeitsdimensionen Neurotizismus, Extraversion, Offenheit, Verträglichkeit und Gewissenhaftigkeit (z.B. Goldberg, 1990; McCrae & Costa, 1985).

Ein Kompetenzkonstrukt definiert sich primär aus relevanten Situationen und Anforderungen, die zugrunde liegende Faktorenstruktur muss weder eindimensional sein noch eine Einfachstruktur aufweisen (s.u., psychometrische Modelle). Eine induktive, primär faktorenanalytisch gesteuerte Auswahl von Aufgaben dürfte daher nur in den wenigsten Bereichen der Kompetenzdiagnostik zweckmäßig sein.

Kriteriumssampling

Eine in der psychologischen Diagnostik weniger häufig verfolgte Vorgehensweise zur Erstellung diagnostischer Instrumente ist das Sampling von Verhaltensweisen aus dem interessierenden Bereich – das Kriteriumssampling (engl. *criterion sampling*). Hierbei wird aus einem interessierenden Verhaltensbereich, in dem erfolgreiches Handeln vorhergesagt werden soll, eine möglichst repräsentative Stichprobe von Verhaltensweisen ausgewählt. Diese werden dann in Testaufgaben überführt. Eine in der Personalauswahl angewandte Methode, die die Idee des Kriteriumssampling umsetzt, sind Assessment Center (z.B. Schuler & Stehle, 1992). Wenngleich diese Methode sich bei der Vorhersage von Berufserfolg bewährt hat (z.B. Gaugler, Rosenthal, Thornton & Bentson, 1987), ist sie für viele – v.a. wissenschaftliche – Untersuchungskontexte, bei denen größere Stichproben untersucht werden sollen, zu aufwendig.

Kriteriumssampling ist ein für Kompetenzdiagnostik sehr nahe liegendes Vorgehen und wird von McClelland (1973) als die optimale Vorgehensweise zur Entwicklung von Kompetenztests betrachtet. Tatsächlich ist die Zweckmäßigkeit dieser Methode, wenn man Kompetenzen als kontextspezifische, durch einen bestimmten Situationsbereich definierte Konstrukte betrachtet, bestechend. Es sollte jedoch nicht übersehen werden, dass damit auch spezifische Schwierigkeiten verbunden sind, so müssen z.B. transparente Kriterien für eine „repräsentative" Auswahl von Verhaltensweisen definiert werden. Eine große Herausforderung bei der Entwicklung eines ökonomischen Messverfahrens stellt zudem die Konstruktion von konkreten Aufgaben dar, die einfach zu bearbeiten

und auszuwerten sind und dennoch die interessierenden Verhaltensweisen angemessen repräsentieren.

Zusammenfassung

Eine externale Konstruktion oder ein induktives Vorgehen erscheinen für die Diagnostik von Kompetenzen wenig zweckmäßig. Vor allem im Bezug auf das in der Psychologie verbreitete faktorenanalytische Vorgehen ist nochmals darauf hinzuweisen, dass eine Selektion von Aufgaben immer auch eine Änderung des gemessenen Konstruktes bedeutet. Im Rahmen der Diagnostik von theoretisch als eindimensional angenommenen Konstrukten mag dies angemessen sein. Der Ausschluss eines Items mit einer niedrigen Faktorladung erfolgt unter der Prämisse, dass diese Faktorladung darauf hinweise, dass das Item „nicht das (selbe) Konstrukt misst". Wenn hingegen für einen Kompetenztest Aufgaben ausgewählt werden, die als relevant für einen bestimmten Situationsbereich betrachtet werden, ist diese Argumentation nicht angemessen. Items, die mit den übrigen nicht oder nur niedrig korrelieren, repräsentieren immer noch Verhaltensweisen aus dem interessierenden Kontext, und ein Ausschluss derselben aus dem Messinstrument führt zu einer – nicht inhaltlich oder theoretisch begründeten – Reduktion des ursprünglichen Konstrukts.

Zusammenfassend stellen sowohl eine deduktive Aufgabenauswahl als auch das Kriteriumssampling viel versprechende Testentwicklungsmethoden im Bereich der Kompetenzdiagnostik dar. Aus beiden Vorgehensweisen können fundierte Begründungen für bestimmte Aufgaben abgeleitet werden. Die Argumentation unterscheidet sich in beiden Fällen, was sich auch in Unterschieden in den entwickelten Aufgaben niederschlagen kann.

Bei einer *deduktiven Vorgehensweise* wird zunächst ein Kompetenzkonstrukt beschrieben - idealerweise mit einem theoretischen Modell, das sowohl Personen- als auch Situationsmerkmale einbezieht. Aus diesem Modell lassen sich Aufgaben ableiten, die von kompetenten Personen mit einer höheren Wahrscheinlichkeit gelöst werden können als von einer weniger kompetenten Person. Diese theoretisch fundierte Annahme über den Zusammenhang zwischen der interessierenden Kompetenz und der Aufgabenlösung stellt beim deduktiven Vorgehen die Begründung für die Verwendung eines Aufgabeninhaltes dar.

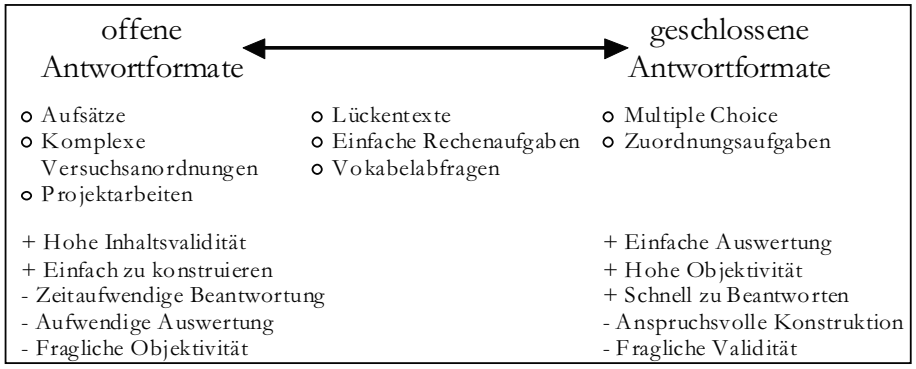
Im Falle des *Kriteriumssamplings* wird der für eine Kompetenz relevante Situations- und Verhaltensbereich systematisch beschrieben. Die Begründung für einen Aufgabeninhalt basiert darauf, dass die Aufgabe eine inhaltliche Entsprechung zu realen Situationen im interessierenden Bereich darstellt.

2.3.3 Antwortformate

Der Inhalt einer Testaufgabe besteht zum einen aus einem Stimulus, in dem ein zu lösendes Problem präsentiert wird – z.B. ein zu lesender Text in Kombination mit auf den Text bezogenen Fragen. Der zweite wesentliche Teil besteht aus den Vorgaben, die der getesteten Person für ihre Antwort gemacht werden. Eine insbesondere in der psychologischen Diagnostik verwendete Möglichkeit sind *geschlossenen Antwortformate*, bei denen mehrere Lösungen vorgegeben werden, aus denen die getestete Person eine oder mehrere auswählen muss (*multiple choice*). Eine Aufgabe ist gelöst, wenn die richtigen Antwortalternativen gewählt

werden. Alternativ kann eine Reihe von Lösungen für eine Anzahl Probleme vorgegeben werden (z.B. Fremdsprachvokabeln mit Übersetzungen), die einander zugeordnet werden müssen. Einen Gegenpol bilden *offene Antwortformate*, bei denen die getestete Person eine Antwort bzw. Lösung frei produziert und in der Regel schriftlich dokumentiert. Die Korrektheit oder Qualität der Lösung muss in diesen Fällen von Beurteilern festgestellt werden, hierfür wird wiederum eine sorgfältige Dokumentation der Beurteilungskriterien benötigt. Geschlossene und offene Antworten können als Pole eines Kontinuums betrachtet werden, das von einer vollständigen Vorgabe von Antwortalternativen über relativ kurze offene Antworten (z.B. Lückentexte oder einfache Rechenaufgaben) bis zu komplexen Produkten wie Aufsätze oder Briefe zu einem gegebenen Thema, die Konstruktion von naturwissenschaftlichen Versuchsanordnungen oder das gemeinsame Entwickeln von Projekten reicht.

Abbildung 2.2: Kontinuum offener und geschlossener Antwortformate für Kompetenztests mit Beispielen sowie Vor- und Nachteilen.



Die Vorteile geschlossener Antworten sind eine hohe zeitliche Ökonomie bei Beantwortung und Auswertung sowie eine hohe Auswertungsobjektivität. Ein Nachteil ist, dass viele Konstrukte einer Messung mit Multiple-Choice-Aufgaben nur schwer zugänglich sind, da das Erkennen einer richtigen Antwort ein anderer und in der Regel einfacherer Prozess ist als das eigenständige Produzieren einer Lösung. Bei der Erfassung von Kompetenzen stellt sich die Frage, ob von der künstlichen Situation vorgegebener Antwortkategorien auf die Bewältigung realer Situationen geschlossen werden kann, d.h. ob eine derartige Messung valide ist. Um ein theoretisches Konstrukt in eine Aufgabe mit geschlossenem Antwortformat zu übersetzen, ist zudem eine gründliche theoretische Vorarbeit notwendig. Aufgaben mit offenem Antwortformat sind hingegen zunächst leichter zu konstruieren, und das in offenen Antworten gezeigte, komplexere Verhalten lässt sich leichter auf reale Situationen übertragen. Die Nachteile offener Antworten, insbesondere mit einem hohen Freiheitsgrad für die getestete Person, liegen in einem relativ hohen zeitlichen Aufwand für die Beantwortung sowie einer aufwendigen Auswertung mit einer unter Umständen fraglichen Objektivität. Abbildung 2.2 gibt einen zusammenfassenden Überblick der Vor- und Nachteile offener und geschlossener Antwortformate.

Im Hinblick auf technologiebasiertes Assessment von Kompetenzen ist zu den möglichen Antwortformaten festzustellen, dass vor allem geschlossene Antwortformate direkt in computerisierte Testversionen übertragbar sind. Ob eine Antwort durch eine schriftliche Markierung oder per Maus gewählt wird, dürfte das Antwortverhalten kaum beeinflussen (s. auch Kapitel 3). Im Hinblick auf offene Antwortformate lassen sich größere Effekte einer Computerisierung erwarten, das Produzieren eines schriftlichen Textes am Computer stellt z.B. andere Anforderungen als eine handschriftliche Anfertigung. Hier ist sorgfältig zu prüfen, inwieweit technologiebasierte Messinstrumente die gleichen Konstrukte messen wie handschriftliche Tests. Andererseits entstehen bei der Kompetenzmessung am Computer auch neue Möglichkeiten, flexible Antwortformate zu gestalten, die ohne Computereinsatz nicht realisierbar währen, z.B. die Interaktion mit dynamischen Systemen (z.B. Kröner, 2001; Wirth, 2004). Auf neue Möglichkeiten der Aufgabengestaltung am Computer wird in Kapitel 4 dieses Bandes ausführlicher eingegangen.

2.3.4 Schwierigkeitsbestimmende Aufgabenmerkmale

Da bei der Definition von Kompetenzen ausgewählte Situationen und die durch diese gestellten Anforderungen im Mittelpunkt stehen, ist es für die Diagnostik und Modellierung von Kompetenzen von besonderem Interesse, diese Situationen systematisch zu beschreiben. Interessant ist dabei vor allem, welche Eigenschaften einer Situation die erfolgreiche Bewältigung für die handelnden Personen erschweren oder erleichtern. Das Wissen über situative Charakteristika, die erfolgreiches Handeln mit bestimmen, ermöglicht ein tieferes Verständnis der Prozesse, die dem erfolgreichen Handeln zugrunde liegen, und damit ein besseres Verständnis des interessierenden Kompetenzkonstrukts.

Im Kontext der Testentwicklung ist die Beschreibung relevanter Situationsmerkmale in eine Beschreibung der zur Erfassung eines Kompetenzkonstrukts eingesetzten Aufgaben zu übertragen; die Aufgaben werden hinsichtlich *schwierigkeitsbestimmender Merkmale* beschrieben. Diese beziehen sich auf Eigenschaften der Testaufgaben, die mit höheren oder niedrigeren Anforderungen an die getesteten Personen einhergehen und damit die Lösungswahrscheinlichkeiten der Aufgaben beeinflussen. Die Definition und Zuordnung solcher Aufgabenmerkmale setzt theoretische Annahmen darüber voraus, welche Prozesse bei der Aufgabenbearbeitung ablaufen und wie diese durch situative Merkmale beeinflusst werden. Die Prüfung dieser Annahmen durch die Vorhersage von Aufgabenschwierigkeiten kann als eine Prüfung der Validität des Messinstrumentes betrachtet werden (Borsboom et al., 2004; Embretson, 1983, 1998).

Eine weitere, für die Diagnostik von Kompetenzen interessante Funktion schwierigkeitsbestimmender Aufgabenmerkmale ist ihre mögliche Verwendung bei der Definition von Kompetenzniveaus. Wenn es gelingt, die beobachteten Unterschiede in der Schwierigkeit von Testaufgaben durch eine Reihe von Aufgabenmerkmalen zu erklären, so können diese Merkmale herangezogen werden, unterschiedliche Niveaus des gemessenen Konstruktes zu beschreiben. Dies ermöglicht eine über die konkreten eingesetzten Testaufgaben hinaus generalisierte Niveaudefinition, die zugleich empirisch fundiert ist (vgl. Hartig, 2007). Anwendungsbeispiele für die Verwendung von Aufgabenmerkmalen zur Be-

schreibung von Kompetenzskalen finden sich in der DESI-Studie zur Erfassung von Sprachkompetenzen im Englischen (Beck & Klieme, 2007), aber auch in der Berufspädagogik (Seeber, 2005). Insgesamt erscheint es ausgesprochen lohnend, die systematische Beschreibung von Situationen bzw. Testaufgaben hinsichtlich schwierigkeitsbestimmender Merkmale sowohl bei der Definition und Beschreibung eines Kompetenzkonstrukts als auch bei der Testentwicklung frühzeitig zu berücksichtigen.

2.4 Psychometrische Modelle zur Messung von Kompetenzen

Ziel der Auswertung eines pädagogisch-psychologischen Tests ist es, quantifizierende und/oder klassifizierende Aussagen über interessierende Merkmalsausprägungen der getesteten Personen zu machen. Dies geschieht z.B. im einfachsten Fall durch das Zusammenfassen der Antworten auf eine Menge von Aufgaben zu einem einzelnen Testwert, hinsichtlich dessen Personen als „mehr" oder „weniger" kompetent beschrieben werden. Derartigen Auswertungsroutinen liegen implizit oder explizit bestimmte Annahmen über Zusammenhänge zwischen dem zu messenden Merkmal und dem beobachteten Testverhalten zugrunde. Wenn z.B. alle Aufgaben zu einem gemeinsamen Testwert zusammengefasst werden, beruht das auf den Annahmen, dass (a) das zu messende Merkmal als ein einzelnes quantitatives Kontinuum beschrieben werden kann und (b) das Lösen oder Nichtlösen aller Aufgaben auf interindividuelle Unterschiede in diesem Merkmal zurückgeht.

Annahmen wie in diesem einfachen Beispiel bilden ein *psychometrisches Modell*. Die Funktion eines psychometrischen Modells ist zum einen, die angenommenen Zusammenhänge zwischen Merkmal und Testverhalten zu beschreiben. Aus dem Modell werden zum anderen, in Einklang mit diesen Annahmen, die Auswertungsroutinen abgeleitet, mit denen auf Basis des beobachteten Testverhaltens die individuellen Testwerte für die getesteten Personen ermittelt werden. Im Kontext psychometrischer Modelle werden die interessierenden Merkmale häufig als so genannte *latente Variablen* bezeichnet, die nicht unmittelbar beobachtet werden können, sondern über das Testverhalten – die *beobachteten Variablen* oder *Indikatoren* – erschlossen werden müssen. Die Messung der interessierenden Merkmale besteht darin, auf Basis der Werte einer Person in den beobachteten Variablen und unter Annahme der im psychometrischen Modell formulierten Zusammenhänge Schätzungen für die latenten Variablen vorzunehmen.

Das beobachtete Verhalten in einem Test wird in der Regel in bestimmten Kategorien klassifiziert – z.B. in „richtige" und „falsche" Antworten. Für derart kategorisierte Antworten werden psychometrische Modelle in Form von Wahrscheinlichkeitsaussagen formuliert; z.B. wie ändert sich die Wahrscheinlichkeit einer richtigen Antwort in Abhängigkeit von der latenten (Fähigkeits-)Variable. Die Beziehungen zwischen Merkmalen und Testverhalten werden in psychometrischen Modellen in Form mathematischer Funktionen beschrieben. Ein großer Teil dieser psychometrischen Modelle, die sich mit den Antworten (engl. *responses*) auf einzelne *Items* befassen, wird unter dem bereits erwähnten Begriff der *Item-Response-Theorie* (IRT) zusammengefasst. Da in diesen Modellen Wahr-

scheinlichkeitsaussagen formuliert werden, werden auch die Bezeichnungen *probabilistische Testtheorie* oder *probabilistische Testmodelle* verwendet.

2.4.1 Kompetenzdimensionen und Kompetenzstrukturen

Die am häufigsten angewandten Modelle der IRT beinhalten zur Modellierung von Unterschieden zwischen Personen eine einzelne, kontinuierlich latente Variable (z.B. Embretson & Reise, 2000; Wilson, 2005), d.h. es handelt sich um *eindimensionale Modelle*. Inhaltlich bedeutet die Anwendung derartiger Modelle in der Kompetenzdiagnostik, dass Unterschiede in der zu erfassenden Kompetenz auf einem einzelnen Kontinuum beschrieben werden. Die eindimensionale Modellierung und Erfassung von Kompetenzkonstrukten überwiegt derzeit in der pädagogisch-psychologischen Diagnostik. So wurden z.B. Schülerkompetenzen in den großen Schulleistungsstudien der letzen Jahre wie PISA, PIRLS, TIMSS oder DESI im Wesentlichen mit Tests und psychometrischen Modellen gemessen, bei denen jede erfasste Kompetenz einer einzelnen latenten Dimension entspricht (vgl. Adams, 2005; Adams & Wu, 2002; Gonzalez, 2003; Gonzalez, Galia & Li, 2004; Hartig, Jude & Wagner, in Druck). Eindimensionale Modelle haben den Vorteil, dass Analyse und Ergebnisinterpretation relativ einfach sind. Gleichzeitig ist jedoch auch das dabei angenommene Kompetenzmodell ein sehr einfaches. Alle Personen und alle Aufgaben werden auf einem einzelnen Kontinuum von niedriger bis hoher Kompetenz bzw. niedriger bis hoher Aufgabenschwierigkeit angeordnet.

Um differenziertere Aussagen über die Kompetenzen und Teilkompetenzen der untersuchten Personen zu machen und um komplexeren Kompetenzkonstrukten gerecht zu werden, können auch mehrdimensionale IRT-Modelle verwendet werden. Mehrdimensionale Modelle enthalten mehrere latente Variablen, mit denen die in einem Test erfasste Kompetenz differenziert hinsichtlich mehrerer zugrunde liegender Teilkompetenzen – z.B. spezifische Fertigkeiten oder spezifisches Wissen – modelliert werden kann (z.B. Ackerman, Gierl, & Walker, 2003). Anwendung haben mehrdimensionale IRT-Modelle z.B. im Kontext der PISA-Studien gefunden, wo teilweise mehrdimensionale Modellierungen innerhalb einzelner Kompetenzbereiche vorgenommen wurden. In den jeweiligen Schwerpunkt-Bereichen (Lesen in PISA 2000, Mathematik in PISA 2003 und Naturwissenschaften in PISA 2006) erfolgte eine differenzierte mehrdimensionale Auswertung, bei der Teilkompetenzen als separate Dimensionen modelliert wurden. Auf diese Weise wurden zum Beispiel für PISA 2003 nicht nur für mathematische Kompetenz insgesamt, sondern auch getrennt für vier „übergreifende Ideen“ Kompetenzschätzungen vorgenommen (vgl. Blum et al., 2004).

Mehrdimensionale IRT-Modelle können bei der Auswertung von Tests auch verwendet werden, um Zusammenhänge zwischen verschiedenen Kompetenzen zu untersuchen, die jeweils durch eine separate latente Dimension modelliert werden. Auf diese Weise können Zusammenhänge zwischen verschiedenen Kompetenzkonstrukten besser geschätzt werden. Einzelne Kompetenzkonstrukte werden jedoch weiterhin durch jeweils ein einzelnes Kontinuum abgebildet. Derartige mehrdimensionale IRT-Modelle kamen in großem Maßstab ebenfalls in den PISA-Studien zur Anwendung, indem die Daten für die

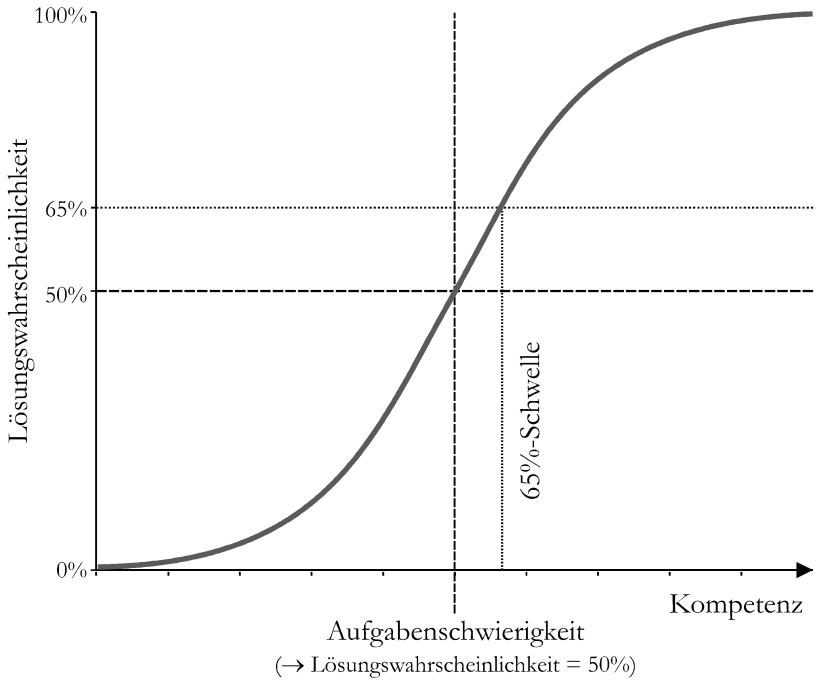
drei Bereiche Lesekompetenz, mathematische Kompetenz und naturwissenschaftliche Kompetenz (2003 auch Problemlösekompetenz) gemeinsam in einem mehrdimensionalen Modell analysiert wurden (Adams, 2005; Adams & Wu, 2002). In DESI (vgl. Beck & Klieme, 2007) wurde ein mehrdimensionales IRT-Modell verwendet, um für einzelne Kompetenzen zu zwei Messzeitpunkten zu modellieren (Hartig & Kühnbach, 2006).

2.4.2 Modellierung von Kompetenzniveaus

Im Falle einer eindimensionalen Modellierung wird eine zu erfassende Kompetenz auf einem einzelnen Kontinuum abgebildet, das ohne Abstufung von einer niedrigen über eine durchschnittliche bis zu einer hohen Kompetenz reicht. Das Anliegen von Kompetenzniveaumodellen ist es, derartige kontinuierliche Dimensionen inhaltlich konkret zu beschreiben, d.h. die Werte auf den kontinuierlichen Skalen in Bezug zu spezifischen, im Test realisierten Anforderungen zu setzen (z.B. Beaton & Allen, 1992; Hartig, 2007; Hartig & Klieme, 2006; Wilson, 2005). Die „Kompetenzstufen", anhand derer die Ergebnisse der PISA-Studien aufbereitet und berichtet wurden, sind ein prominentes Beispiel für eine derartige Beschreibung von Kompetenzniveaus (z.B. OECD, 2001, 2004a,b). Durch die Definition von Niveaus wird es möglich, die Kompetenzen untersuchter Personen und Personengruppen nicht nur bezogen auf soziale Bezugsnormen zu beschreiben. Indem eine Aussage darüber gemacht wird, welche spezifischen Anforderungen z.B. Schüler auf einem bestimmten Kompetenzniveau bewältigen können und welche nicht, kann eine *kriterienorientierte* Testwertinterpretation vorgenommen werden (vgl. Kapitel 1).

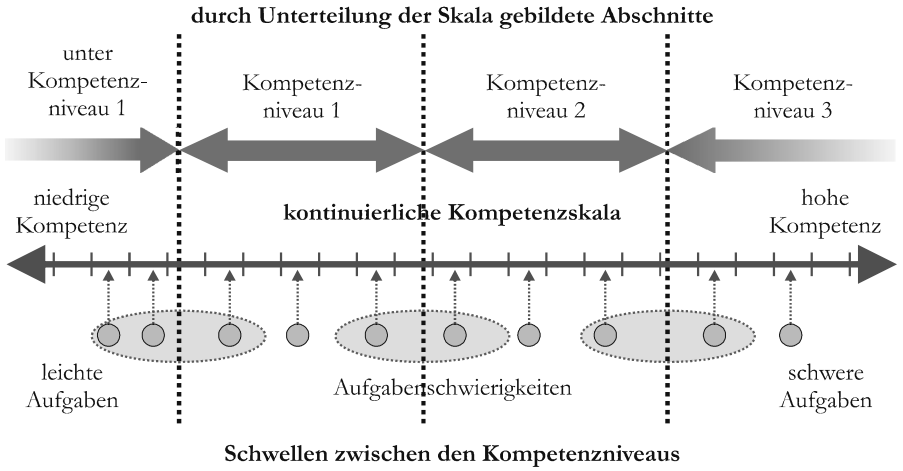
Moderne Methoden zur Definition von Kompetenzniveaus machen von Techniken der IRT Gebrauch, da mit diesen gemeinsame Skalen für die Kompetenzen der getesteten Personen und die Schwierigkeiten der eingesetzten Testaufgaben konstruiert werden können. Diese gemeinsame Skala wird dadurch konstruiert, dass Zusammenhänge zwischen der zu messenden Kompetenz und den Antwortwahrscheinlichkeiten für die verwendeten Aufgaben formuliert werden. Ein derartiger Zusammenhang wird in Abbildung 2.3 am Beispiel eines der einfachsten IRT-Modelle, dem dichotomen Raschmodell (Rasch, 1960), illustriert. Die Schwierigkeit einer Aufgabe wird in diesem Modell dort verortet, wo Personen mit einer entsprechend hohen Kompetenz die Aufgabe zu 50% lösen. Durch die so konstruierte Kompetenzskala können die Ausprägungen der gemessenen Kompetenz zu den Lösungswahrscheinlichkeiten der verwendeten Aufgaben in Beziehung gesetzt werden. Der angenommene funktionale Zusammenhang zwischen Kompetenz und Lösungswahrscheinlichkeit erlaubt es hierbei, Punkte auf der Kompetenzskala zu bestimmen, an denen die Lösungswahrscheinlichkeit einen höheren Wert als 50% annimmt. So wurden z.B. die zur Beschreibung der Kompetenzskalen in DESI diejenigen Punkte verwendet, an denen die Lösungswahrscheinlichkeiten der Aufgaben 65% betragen. Die Bildung dieser „65%-Schwelle" auf Basis des funktionalen Zusammenhangs zwischen Kompetenz und Lösungswahrscheinlichkeit ist in der folgenden Abbildung veranschaulicht.

Abbildung 2.3: Verortung der Schwierigkeit einer einzelnen Testaufgabe auf einer gemeinsamen Skala mit den Ausprägungen einer zu messenden Kompetenz am Beispiel des dichotomen Rasch-Modells.



Bei der Definition von Kompetenzniveaus ist nun die entscheidende Frage, wo die *Grenzen* zwischen den Niveaus gezogen werden. Innerhalb eines Skalenabschnittes, der als ein Niveau betrachtet wird, wird dann keine weitere inhaltliche Differenzierung der erfassten Kompetenz vorgenommen. Die inhaltliche Beschreibung der Kompetenz erfolgt nur für jeden der gebildeten Skalenabschnitte. Um zu diesen Schwellen zwischen den Niveaus zu gelangen, wird von der gemeinsamen Abbildung der gemessenen Kompetenz und der Aufgabenschwierigkeiten Gebrauch gemacht. Für die inhaltliche Charakterisierung eines Kompetenzniveaus sind diejenigen Aufgaben relevant, deren Schwierigkeiten auf oder in der Nähe der unteren Schwelle dieses Niveaus liegen. Es sind diese Aufgaben, die von Personen auf diesem Niveau mit hinreichender Wahrscheinlichkeit bewältigt werden können und von den Personen auf dem Niveau unterhalb davon noch nicht. Die Beziehung zwischen Aufgabenschwierigkeiten und Schwellen zwischen den Niveaus auf einer kontinuierlichen Kompetenzskala ist in Abbildung 2.4 schematisch veranschaulicht.

Abbildung 2.4: Veranschaulichung der Unterteilung einer kontinuierlichen Kompetenzskala mit darauf verorteten Aufgabenschwierigkeiten in Kompetenzniveaus. Die Aufgabenschwierigkeiten in Nachbarschaft der Schwellen zwischen den Niveaus sind unterlegt (nach Hartig, 2007).



Um die Schwellen und die inhaltliche Beschreibung der Kompetenzniveaus zu definieren, sind in verschiedenen Studien unterschiedlich stark modellgeleitete Vorgehen gewählt worden. Im einfachsten Fall werden die Inhalte der Testaufgaben, die ähnliche Schwierigkeiten aufweisen, post hoc ohne bestimmte Vorannahmen interpretiert. Die inhaltliche Beschreibung der Skalenabschnitte erfolgt dann anhand der Ergebnisse dieser Post-Hoc-Analyse der Aufgabeninhalte. Die Schwellen zwischen den Niveaus lassen sich jedoch auch über modellgeleitete Vorgehen definieren. Diese ermöglichen es, Kompetenzniveaus nicht nur durch einzelne Aufgaben zu definieren, sondern sich auf generalisierbare Anforderungen zu beziehen, welche die interessierenden Kompetenzen definieren. Für ein derartiges stärker theoriegeleitetes Vorgehen ist es notwendig, die Testaufgaben schon vorab hinsichtlich ihrer Anforderungen zu beschreiben. Beschreibungen möglicher modellgeleiteter Vorgehen zur Definition von Kompetenzniveaus finden sich z.B. bei Hartig (2007) oder Janssen, Tuerlinckx, Meulders & De Boeck (2000).

3 Computer- und netzwerkbasiertes Assessment

Astrid Jurecka & Johannes Hartig

Im Bereich des technologiebasierten Assessments wird eine große Anzahl von Begriffen verwendet, die aus der englischsprachigen Literatur übernommenen worden sind. Da einige dieser Begriffe synonym sind oder deutliche Überschneidungen aufweisen, soll in den folgenden Abschnitten eine kurze Erläuterung der relevanten Termini erfolgen sowie der Gebrauch im Kontext der vorliegenden Expertise definiert werden. Anschließend werden im vorliegenden Kapitel generelle kritische Aspekte und Vorteile computerbasierten Assessments erläutert, der letzte Abschnitt behandelt spezifische Vor- und Nachteile der Nutzung von Computernetzwerken für diagnostische Anwendungen.

3.1 Zentrale Begriffe des computer- und netzwerkbasierten Assessments

Die Verwendungsmöglichkeiten von Computern im psychologisch- oder pädagogisch-diagnostischen Bereich sind vielfältig. Neben der Vorgabe von in der Regel standardisierten Tests sind z.B. auch computerunterstützte Interviews (z.B. *Computer Assisted Telephone Interviews*, CATI) oder Programme zur computerunterstützten Testinterpretation zu nennen. Im vorliegenden Kapitel wird von computerbasiertem Assessment als übergeordnetem Sammelbegriff gesprochen, der auch halb- und unstandardisierte Beurteilungsmethoden einschließt. Die meisten existierenden Technologie-Anwendungen in der pädagogisch-psychologischen Diagnostik erfüllen jedoch auch die engeren Kriterien standardisierter Testverfahren (vgl. Kapitel 2).

3.1.1 Technologiebasiertes Assessment (TBA)

Englisch: Technology Based Assessment (TBA); synonym: electronic assessment, e-assessment.

Unter dem sehr allgemeinen Begriff *technologiebasiertes Assessment (TBA)* lässt sich jegliche Verwendung von Informationstechnologie (IT) in der psychologischen oder pädagogischen Diagnostik zusammenfassen. Wenngleich die hierbei verwendete Technologie in der Regel den Einsatz von Computern im Sinne von z.B. PCs oder Notebooks beinhaltet, muss das nicht unbedingt der Fall sein. So kommen z.B. auch kleinere mobile Geräte (Hand-Held-Computer, Personal Digital Assistants [PDAs]) zum Einsatz. Diese sind vor allem dann von Vorteil, wenn Personen über einen längeren Zeitraum und in ihrem natürlichen Tagesablauf befragt werden sollen. PDAs können hierbei verwendet werden, um den Untersuchungsteilnehmer zu zufälligen Zeitpunkten aufzufordern, bestimmte Fragen zu beantworten, die Antworten werden dann für eine spätere Auswertung gespeichert. Auch der Einsatz von Festnetz- oder Mobiltelefonen zu Befra-

gungszwecken fällt im weiteren Sinne unter TBA. Als weitgehend synonym zu TBA kann der Begriff des *electronic assessment* oder *e-assessment* betrachtet werden.

3.1.2 Computerbasiertes Assessment (CBA) und computerbasiertes Testen (CBT)

Englisch: Computer-Based Assessment, Computer-Based Testing; synonym: Computergestütztes Testen.

Computerbasiertes Assessment bezeichnet im vorliegenden Kontext generell die Verwendung von Computern zur Erfassung psychologischer Merkmale und Konstrukte. Hierbei erfolgt zumindest die Vorgabe des Testmaterials am Computerbildschirm (on-screen presentation), und die getestete Person reagiert ebenfalls über den Computer (z.B. Tastatur oder Maus). Die einzelnen Reaktionen werden zumindest elektronisch aufgezeichnet. In der Regel wird jedoch auch die Testauswertung und ggf. Ergebnisrückmeldung computergesteuert vorgenommen. Computerbasiertes Assessment bezieht sich sowohl auf die Vorgabe diagnostischer Anwendungen an einem einzelnen PC, als auch auf netzwerk- und internetbasiertes Assessment (s.u.). Bei der Vorgabe von Tests auf einem einzelnen PC wird in der Regel entsprechende Testsoftware, die zur Beantwortung durch einen Probanden gestartet werden muss, lokal auf der Festplatte installiert.

Die meisten derzeit existierenden computerbasierten Tests sind standardisierte Tests, die aus einzelnen Aufgaben mit vorgegebenen zugehörigen Antwortalternativen bestehen und große Ähnlichkeit zu auf Papier gedruckten Tests haben (s.u.). Die Testvorgabe am Computer ermöglicht jedoch auch andere Arten von Testmaterial und Reaktionsweisen, wie z.B. Computersimulationen komplexer Systeme (z.B. Brehmer & Dörner, 1993; Dörner & Wearing, 1995; Kröner, 2001) oder virtuelle Realitäten (*virtual realities*), in denen ein Proband sich bewegen und handeln kann (z.B. Frey, Hartig, Ketzel, Zinkernagel & Moosbrugger, im Druck; Hartig, Frey & Ketzel, 2003;).

3.1.3 Papier-Bleistift-Tests

Englisch: Paper-Pencil Tests, PPT

Mit *Papier-Bleistift-Tests* wird im Kontrast zu computerbasierten Tests die übliche Vorgabe des Testmaterials auf einem gedruckten Medium (z.B. Fragebogen, Testheft) und einer schriftlichen Beantwortung bezeichnet.

3.1.4 Computerisiertes Adaptives Testen (CAT)

Englisch: Computer-Adaptive Testing, Computerized Adaptive Testing, Tailored Testing; synonym: Computer-adaptives Testen, Adaptive Computertests.

Computerisiertes Adaptives Testen (CAT) dürfte zu den ältesten und häufigsten Methoden in der psychologisch-pädagogischen Diagnostik gehören, die ohne Computer nicht realisierbar sind. Bei der Vorgabe von Testverfahren, bei denen allen Probanden die gleichen Aufgaben präsentiert werden, bearbeiten in der Regel viele Personen Aufgaben, die für sie deutlich zu leicht oder deutlich zu schwer sind. Adaptives Testen zielt darauf ab, die Schwierigkeiten der Aufga-

ben an die individuelle Fähigkeit des Probanden anzupassen (*tailored testing*; Gershon, 2005). Hierdurch soll die Unter- und Überforderung von Probanden vermieden werden. Zugleich kann mit einer geringeren Anzahl von Aufgaben für den Messwert einer Person die gleiche Messgenauigkeit erreicht werden wie mit einer festen Auswahl von Aufgaben, so dass adaptives Testen eine Erhöhung der Testökonomie verspricht. Als weiterer Vorteil adaptiven Testens werden auch Effekte auf die Motivation der getesteten Personen diskutiert (z.B. Frey, 2006).

Adaptives Testen setzt in der Regel eine Auswertung mit Modellen der Item-Response-Theorie (IRT, vgl. Kap. 2) voraus, da Testergebnisse, die auf unterschiedlichen Aufgabenmengen basieren, auf einer gemeinsamen Skala miteinander verglichen werden sollen. Im Verlauf eines adaptiven Tests wird typischerweise nach jeder Aufgabenbeantwortung eine Schätzung der individuellen Merkmalsausprägung vorgenommen. Auf Basis dieser Schätzung wird dann die nächste Aufgaben dahingehend ausgewählt, dass sie der Fähigkeit der getesteten Person möglichst gut entspricht und daher maximal viel Information über die jeweilige Person liefert (vgl. z.B. Eggen, im Druck; van der Linden & Glas, 2000).

Die Schätzung der individuellen Merkmalsausprägung und die darauf basierende Aufgabenauswahl sind faktisch nur mit Computerunterstützung realisierbar. Zwar gibt es auch Ansätze, bei Papier-Bleistift-Tests in einem gestuften Vorgehen eine zwischengeschaltete Auswertung vorzunehmen und darauf folgende Testteile auf Basis des Ergebnisses auszuwählen (*branched testing*; *verzweigtes Testen*). Verglichen mit der Verbreitung Computerisiertem Adaptiven Testens sind diese Ansätze jedoch vernachlässigbar, adaptives Testen ist in Forschung und Praxis in der Regel gleichbedeutend mit Computerisiertem Adaptiven Testen.

3.1.5 Netzwerkbasiertes Assessment und netzwerkbasiertes Testen (NBA)

Englisch: network-based assessment.

Netzwerkbasiertes Assessment bezeichnet die Verwendung und Vorgabe von Tests innerhalb eines Computernetzwerks. Hierbei kann es sich um ein lokales Netzwerk (Local Area Network, LAN), das Internet, oder auch um eine Kombination aus beidem handeln.

Eine häufige Anwendung von NBA ist, dass innerhalb eines Netzwerkes an mehreren Computern gleichzeitig Tests oder Testbatterien vorgegeben werden können. Dabei existiert üblicherweise ein Testleiterarbeitsplatz, von dem aus die Testvorgabe gesteuert und später die Daten zusammengeführt und ausgewertet werden können. Vorher werden zumeist sowohl Software als auch Tests auf allen dem Netzwerk angeschlossenen Computern installiert. Die Auswertung findet dann, je nach Testsystem, entweder auf dem jeweiligen Testcomputer, oder aber auf einem zentralen Server statt. Neben einer gesteigerten Ökonomie der Datenerfassung und -auswertung sind in Netzwerkumgebungen jedoch auch interaktive Testformen realisierbar, in denen Paare oder Gruppen von Probanden bestimmte Aufgaben gemeinsam oder im Wettbewerb bearbeiten. Frey, Blunk und Banse (2006) untersuchten z.B. das Bindungsverhalten von Paaren in einer netzwerkbasierten virtuellen Umgebung.

3.1.6 Internetbasiertes Assessment (IBA) und internetbasiertes Testen (IBT)

Englisch: internet-based assessment, web-based assessment, online assessment; synonym: internetbasiertes Assessment, Online-Tests.

Beim internetbasierten Assessment findet der Testprozess ausschließlich via Internet statt. In den meisten Fällen benötigt die beantwortende Person lediglich einen Internet-Zugang sowie dazugehörige Standardsoftware (Browser). Die Auswahl der Aufgaben erfolgt auf einem zentralen Server, auf dem auch die Antworten gespeichert werden. Auch eine Testauswertung kann serverseitig erfolgen, eine Ergebnisrückmeldung wird anschließend wieder an den Browser des Probanden zurückgegeben.

Ein offensichtlicher Vorteil internetbasierter Tests ist die zeitliche und räumliche Unabhängigkeit von Forschenden und Untersuchungsteilnehmern. Die rasante Verbreitung des Internet hat dieses mittlerweile zu einem häufig genutzten Medium für psychologische Tests gemacht. Für viele Forschungsbereiche stellen internetbasierte Untersuchungen eine attraktive Möglichkeit dar, ohne großen Aufwand relativ große Mengen von Untersuchungsteilnehmern zu rekrutieren. Während dies jedoch in Fragebogenuntersuchungen, wie sie z.B. in der Persönlichkeitspsychologie typisch sind, relativ unproblematisch ist, ist die Vorgabe von Leistungstests mit spezifischen Schwierigkeiten verbunden. Die Untersuchungssituation entzieht sich der Kontrolle des Forschenden, d.h. es ist unklar, welche und wie viele Personen den Test unter welchen Rahmenbedingungen bearbeiten. Für Tests, deren Ergebnisse Konsequenzen für die getesteten Personen haben (z.B. in Bewerbungsverfahren), findet IBT dementsprechend kaum Verwendung.

3.1.7 Konventionelle Tests

Englisch: Conventional Tests.

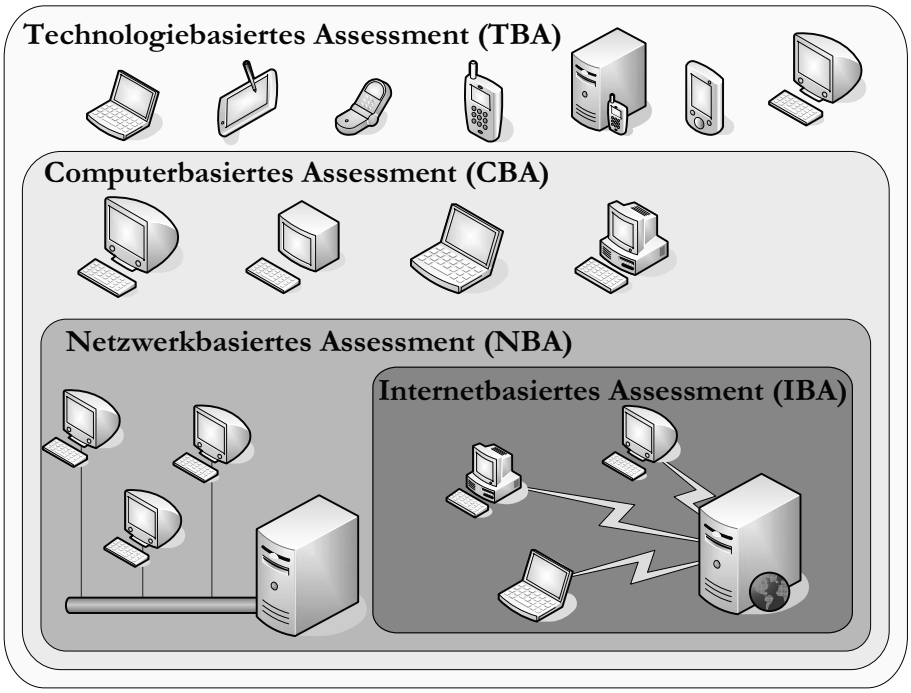
Der Begriff *konventioneller Test* wird häufig als Abgrenzung zu technologiebasierten Testmethoden *verwendet* und ist häufig synonym zu Papier-Bleistift-Tests (s.o.). Der Begriff wird jedoch auch in Abgrenzung zu adaptiven Tests gebraucht, wobei damit Tests bezeichnet werden, bei denen allen Personen dieselben Aufgaben vorgegeben werden – sei es am Computer oder auf Papier. Aufgrund dieser nur aus dem jeweiligen Kontext definierten Bedeutung soll der Begriff konventioneller Test im Folgenden vermieden werden und explizit entweder von Papier-Bleistift-Tests oder nicht-adaptiven Tests gesprochen werden.

3.1.8 Hierarchische Beziehung der verwendeten Begriffe

Die Begriffe Technologiebasiertes Assessment (TBA), Computerbasiertes Assessment (CBA), Netzwerkbasiertes Assessment (NBA) und Internetbasiertes Assessment (IBA) sind im Kontext des vorliegenden Bandes so definiert, dass sie hierarchisch aufeinander aufbauen. TBA ist der breiteste Begriff, CBA ist ein Teilbereich von TBA, in der sich ausschließlich auf die Verwendung von Computern im Sinne von PCs, Notebooks o.ä. bezieht. NBA ist ein Teilbereich von CBT und IBT schließlich ein Teilbereich von NBA. Abbildung 3.1 veran-

schaulicht diese Beziehung. Der vorliegende Band befasst sich im Wesentlichen mit im engeren Sinne computerbasierten Anwendungen zur psychologischen und pädagogischen Diagnostik.

Abbildung 3.1: Übersicht über die hierarchische Beziehung der Begriffe Technologiebasiertes Assessment, Computerbasiertes Assessment, Netzwerkbasiertes Assessment und Internetbasiertes Assessment.



3.2 Kritische Aspekte computerbasierten Assessments

In einer Diskussion computerbasierten Assessments ist es auch notwendig, mögliche negative Effekte und Folgen zu betrachten. Im Folgenden wird auf die Frage nach der Äquivalenz zwischen CBA und Papier-Bleistift-Tests, mögliche Effekte von CBA auf die Testfairness sowie auf ökonomische Aspekte der technischen Voraussetzungen eingegangen.

3.2.1 Äquivalenzproblematik

Durch einen Wechsel des Testmodus, z.B. die Übertragung von Tests von einer Papier-Bleistift-Version in eine computerbasierte Version kann sich unter Umständen die Natur des erfassten Konstruktes verändern, wenn z.B. das Testverhalten durch die computerisierte Darbietung verändert wird. Dies bedeutet dann, dass zwischen den beiden Testversionen keine *Äquivalenz* mehr besteht. Bereits 1986 wurden von der American Psychological Association Richtlinien zum computerbasierten Assessment verfasst. Im Hinblick auf Testäquivalenz konstatieren diese folgendes: „When interpreting scores from the computerized

versions of conventional Tests, the equivalence of scores from computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests." (APA, 1986, S. 18) Die Problematik der *cross-mode equivalence*, d. h. der Vergleichbarkeit und Äquivalenz von Testergebnissen und Testwerten über verschiedene Testmodi hinweg, ist noch immer, auch in neueren Richtlinien zu computerbasiertem Assessment, ein Thema (Harris, 2000; ITC, 2005).

Die Guidelines der International Testing Commission (ITC, 2005) konstatieren, dass Äquivalenz von computerbasierten und Papier-Bleistift-Tests dann gegeben ist, wenn aufgezeigt werden kann, dass beide Versionen (a) vergleichbare Reliabilitäten besitzen, (b) ausreichend hoch miteinander korrelieren, (c) in ähnlicher Größe mit anderen Testverfahren und externen Kriterien korrelieren, und (d) vergleichbare Mittelwerte und Standardabweichungen aufweisen. Ferner sollten beide Versionen ein vergleichbares Level an Flexibilität aufweisen (z.B. Items überarbeiten oder überspringen können), sowie die Präsentation der Items und der Antwortmöglichkeiten auf möglichst ähnliche Art und Weise realisiert sein.

Während die Äquivalenz zwischen computerbasierten und Papier-Bleistift-Versionen von Fragebögen zumeist bestätigt werden kann (z.B. Donovan, Drasgow & Probst, 2000; Hartig, 2003; Rammstedt, Holzinger & Rammsayer, 2004; Rauch, Hartig & Moosbrugger, 2002; Schulenberg & Yutrzenka, 2001), finden sich bei Leistungstests häufiger Effekte der computerbasierten Darbietung auf die Schwierigkeit oder die bei der Beantwortung ablaufenden Prozesse (z.B. Kindsvater & Sturm, 2003; Kobrin & Young, 2003; Pomplun, Frey & Becker, 2002; Troche, Rammstedt & Rammsayer, 2002). Im Laufe der Zeit wurden zahlreiche Studien durchgeführt, die häufig zum Ziel hatten, entweder die Äquivalenz zwischen zwei Testmodi zu bestätigen, oder – im Falle von nicht bestehender Äquivalenz – die Quelle für die Ungleichheit der Testergebnisse zu finden. So führten bereits 1993 Mead und Drasgow eine Meta-Analyse zu Testmodus-Effekten bei Computer- und Papier-Bleistift-Versionen von Power- und Speed-Tests durch. Sie fanden, dass die Korrelation zwischen Computer- und Papier-Bleistift-Verfahren bei Power-Tests $r = .97$ betrug, während die Korrelation bei Speed-Tests lediglich $r = .72$ erreichte. Die Autoren schlussfolgerten daraus, dass die Äquivalenz zwischen Computer- und Papier-Bleistift-Versionen durch einen Speed-Faktor beeinflusst wird. Neuman und Baydoun (1998) hingegen fanden diesbezüglich keine Unterschiede zwischen der auf dem Computer vorgegebenen und der Papier-Bleistift- Version, wenn „speeded CBTs follow the same administration and response procedure as the P&P [Paper & Pencil] format" (S. 71).

Auch andere Wissenschaftler betrachten die Flexibilität der verwendeten Software als eine primäre Quelle der Äquivalenzeinschränkung. Bodman und Robinson (2004) schlussfolgerten im Rahmen mehrerer Experimente, dass Testmodus-Effekte nicht zwingend dadurch zustande kommen, dass es sich um ein Computer- oder um ein Papier-Bleistift-Format handelt, sondern eher von der Flexibilität eines Tests abhängig sind. Flexibilität bedeutet an dieser Stelle beispielsweise, dass es auch bei der computerisierten Form möglich ist, vor- und zurückzublättern sowie ein Item später oder auch erneut zu bearbeiten.

Alles in allem kann wohl konstatiert werden, dass Computer- und Papier-Bleistift-Tests dann vergleichbar sind, wenn bei der Bearbeitung beider Formen dieselben Bedingungen bezüglich der Items und der Testvorgabe gegeben sind.

Zur Problematik der Äquivalenz ist abschließend anzumerken, dass diese selbstverständlich nur relevant ist, wenn (a) ein Test dauerhaft in verschiedenen Modi verwendet werden soll oder (b) bestehende Testnormen und Befunde zur Validität, die auf einer Papier-Bleistift-Version basieren, auf eine computerbasierte Version generalisiert werden sollen oder umgekehrt. In den letzen Jahrzehnten wurden Äquivalenzuntersuchungen vor allem vor dem Hintergrund der letzteren Situation durchgeführt, da mit zunehmender Verbreitung von CBT computerbasierte Fassungen bestehender Papier-Bleistift-Tests erstellt wurden. Der unter (a) genannte Fall ist bei Neuentwicklungen von Tests eher die Ausnahme. Für computerbasierte Tests, die von den spezifischen Möglichkeiten des Mediums Gebrauch machen und dementsprechend nicht in ein gedrucktes Medium übertragbar sind, stellt sich die Äquivalenzfrage nicht.

3.2.2 Technische Voraussetzungen

Wenngleich eine computerbasierte Testdurchführung und -auswertung deutlich ökonomischer sein kann als die Vorgabe von Papier-Bleistift-Tests, darf nicht außer Acht gelassen werden, dass die Bereitstellung der notwendigen Computerhardware und der Testsoftware unter Umständen eine beachtliche ökonomische Belastung darstellen kann. Die möglichen ökonomischen Vorteile von CBA kommen nur dann zum Tragen, wenn zumindest die notwendige Hardware schon für andere Zwecke vorhanden ist oder, falls eine Neuanschaffung erforderlich ist, regelmäßig zu Testzwecken verwendet wird. Die Frage nach den technischen Voraussetzungen stellt sich im Bildungswesen insbesondere in institutionellen Kontexten, in denen größere Probandengruppen (z.B. Schüler oder Studenten) gleichzeitig getestet werden sollen. Auch in Forschungsprojekten mit einer hohen Fallzahl kann die Bereitstellung der notwendigen Hard- und Software in ausreichender Menge ein Argument gegen den Einsatz von CBA darstellen.

3.2.3 Testfairness

Ein im Zusammengang mit CBT häufig diskutierter Kritikpunkt ist die Computer- und Software-Vertrautheit sowohl seitens des Probanden als auch seitens des Testleiters. Die Abhängigkeit von Testergebnissen von der Erfahrung mit dem Computer oder von Computerängstlichkeit kann zu kulturell, ethnisch oder geschlechtlich bedingten Benachteiligungen bei CBT führen, die sich negativ auf das Testergebnis auswirken und zu einer Diskriminierung der betroffenen Personen führen könnten. Es wird also diskutiert, ob CBA die *Testfairness* zuungunsten von Personen beeinträchtigen kann, die im Umgang mit dem Computer weniger vertraut sind. Letztlich ist diese Frage nach der Fairness auch eine Frage der Validität, da dabei implizit angenommen wird, dass interindividuelle Unterschiede in Computererfahrung und Computerängstlichkeit in das Testergebnis eingehen, obwohl diese nicht Gegenstand der Messung sind (z.B. Russel, Goldberg & O'Connor, 2003). Verschiedene Studien zeigen Zu-

sammenhänge von Leistungstestergebnissen mit Computererfahrung (z.B. Lee, 1986) sowie Computerängstlichkeit (z.B. Laguna & Babcock, 1997; Tseng, Tiplady, MacLeod & Wright, 1998). Diese Einflüsse sind jedoch nicht zwingend; in einer Vielzahl von Untersuchungen waren Testleistungen auch unabhängig von Computererfahrung (z.B. Powers & O'Neill, 1993, für einen Überblick siehe Smith, 2003) und Computerängstlichkeit (z.B. Powers & O'Neill, 1993; Wise, Barnes, Harvey & Plake, 1989).

Eine Benachteiligung von computerunerfahrenen Personen kann im Kontext von Diagnostik im Bildungswesen umso schwerwiegender sein, als davon wahrscheinlich insbesondere Gruppen von Personen betroffen sein könnten, die bereits ohnehin schlechte Ausgangsbedingungen haben, wie z.B. Schüler mit einem schwachen sozioökonomischen Hintergrund.

Um mit der Problematik einer möglicherweise reduzierten Testfairness bei CBA umzugehen, können behelfsweise Maße für Computererfahrung (z.B. Richter, Naumann & Groeben, 2001) und Computerängstlichkeit (z.B. Heinssen, Glass & Knight, 1987) mit erhoben werden, um den Einfluss dieser Variablen auf die Testergebnisse abschätzen und ggf. statistisch kontrollieren zu können. In Bereichen, in denen ein deutlicher Effekt einer computerbasierten Testvorgabe auf die Testfairness angenommen werden muss, sollte vor allem dann davon Abstand genommen werden, wenn die Testergebnisse Konsequenzen für Individuen oder Institutionen haben. Es kann aber auch in Betracht gezogen werden, die Problematik durch gezielte Übungen im Umgang mit den computerbasierten Testoberflächen zu reduzieren.

3.3 Vorteile computerbasierten Assessments

Im Folgenden soll nun auf die in der Literatur am häufigsten genannten Vorteile computerbasierten Assessments näher eingegangen werden, die größtenteils auch für netzwerk- und internetbasiertes Assessment gültig sind. Da sich dort aber noch zusätzliche bereichsspezifische Vor- und Nachteile finden lassen, werden netzwerk- und internetbasiertes Assessment im Anschluss in zusätzlichen separaten Abschnitten betrachtet.

3.3.1 Objektivität, Reliabilität und Validität

Zu den am häufigsten genannten Vorteilen computerbasierten Assessments gehört sicherlich die Verbesserung der Testgütekriterien Objektivität, Reliabilität und Validität. Die *Objektivität* eines Tests kann durch die Standardisierung der Testvorgabe, der Testinstruktion und einer standardisierten Auswertungsroutine verbessert werden. Dies führt zu einer Verringerung von Situations- und Testleitereffekten, was wiederum genauere, fehlerfreiere Ergebnisse zur Folge hat. Eine zusätzliche Verringerung möglicher Fehler ist im CBA dadurch gegeben, dass die Aufgabenvorgabe und -bearbeitung besser kontrolliert werden kann und so z.B. besser sichergestellt und kontrolliert werden könnte, dass ein Proband auch tatsächlich alle Items beantwortet oder zumindest vorgegeben bekommen hat. Eine Verringerung von Mess- und Auswertungsfehlern kann schließlich zu einer Verbesserung der *Testreliabilität* (Messgenauigkeit) führen (Ridgeway & McCusker, 2004).

Die Verwendung automatisierter Auswertungsprozesse bereits während des Testens ermöglicht außerdem die Anwendung adaptiver Tests (CAT, s.o.). Diese ermöglichen kürzere, informativere Tests, bestenfalls ohne dabei eine Verringerung der Reliabilität durch eine Verkürzung des Tests befürchten zu müssen. Aufgrund der Möglichkeit, ein im Hinblick auf Schwierigkeit und Information optimales Itemset für einen Probanden auszuwählen kann ferner die *Testvalidität* verbessert werden (z.B. Frey, 2005). Auch die Verwendung vielfältigerer Testmaterialien, die durch CBA möglich wird (s.u.), kann die Validität eines Tests steigern.

3.3.2 Ökonomie von Datenerhebung und -auswertung

Bei einer Aufzählung der Vorteile computerbasierten Assessments sollten auch ökonomische Vorteile nicht außer Acht gelassen werden. Häufig können diese schneller und in kürzerer Zeit vorgegeben werden, was zu einer Zeitersparnis sowohl auf Seiten des Probanden als auch seitens des Testleiters führt. Ferner können durch automatische Testauswertung und automatisches Feedback die Menge des Arbeitsaufwands sowie administrative Kosten verringert werden.

In der Möglichkeit einer sofortigen automatischen Test- und Ergebnisauswertung sowie in der verhältnismäßig einfachen Testadministration liegt auch für den Probanden ein Vorteil, da er nicht mehr dazu gezwungen ist, auf seine Ergebnisse zu warten.

Ein nicht zu unterschätzender Nutzen von CBA ist auch der schlichte Umstand, dass Testdaten zur Auswertung nicht mehr in ein computertaugliches Format übertragen werden müssen, sondern unmittelbar während der Testung aufgezeichnet werden. Hierdurch wird die Dateneingabe und Testauswertung als eine mögliche Fehlerquelle ausgeschlossen, außerdem auch in nicht zu unterschätzendem Maße Arbeitszeit gespart.

Die computerbasierte Testvorgabe und Datenaufzeichnung kann letztlich auch die Verwaltung von Testdaten, etwa bei der Sammlung von Daten an denselben Personen zu mehreren Zeitpunkten, erleichtern.

3.3.3 Vielfältigere Testinhalte und -formate

Eine ganz wesentliche Stärke computerbasierter Testmethoden ist gerade für die Erfassung von Kompetenzen von besonderer Bedeutung: Die Testvorgabe am Computer erlaubt die Verwendung neuer und innovativer Testmaterialien, die bei gedruckten Medien nicht möglich ist. So kann das Testmaterial auch Audio- oder Video-Material enthalten, auch die bereits oben erwähnten Simulationen oder virtuellen Realitäten fallen unter nur am Computer realisierbare Testmaterialien.

Neben der Darbietung vielfältigerer und komplexerer Stimuli ermöglicht CBT auch die Aufzeichnung von Reaktionen, die in Papier-Bleistift-Tests praktisch nicht erfasst werden können, z.B. Bearbeitungs- oder Reaktionszeiten.

Die Möglichkeiten computerbasierter Testmaterialien sind gerade für die Erfassung von Kompetenzen von besonderem Interesse. Aus der Definition von Kompetenzen als situations- und kontextspezifische Konstrukte lassen sich hohe Ansprüche an die Übertragbarkeit von Testergebnissen auf reale Situatio-

nen ableiten. Um diesen Ansprüchen gerecht zu werden, bietet Computertechnologie zur Gestaltung vielfältigerer und komplexerer Testmaterialien neue Möglichkeiten, die bei Papier-Bleistift-Tests nicht vorhanden sind.

In Kapitel 4 dieses Bandes wird ausführlicher auf die Möglichkeiten computerbasierter Testmaterialien eingegangen.

3.3.4 Höhere Lebensnähe

Ein weiteres Argument für die Verwendung computerbasierten Assessments hängt mit der Tatsache zusammen, dass es mittlerweile für eine große Anzahl von Schülern üblich und vollkommen natürlich ist, den Computer im Rahmen schulischer Aktivitäten zu verwenden. Dies kann beispielsweise der Fall sein bei der Erledigung der Hausaufgaben mit Hilfe von Textverarbeitungsprogrammen, bei der Verwendung spezieller Programme zur zusätzlichen Wissensaneignung in bestimmten Fächern oder bei der Verwendung des Internets für Recherchetätigkeiten. Dies führt dazu, dass der Computer mittlerweile für viele Schüler zu einem wichtigen Werkzeug des Lernens und Problemlösens (Ridgway & McCusker, 2004) geworden ist. Daher erscheint es beinahe unangebracht, dass es während der Durchführung von Tests und Prüfungen in Deutschland noch immer unüblich ist, zur Bewältigung der gestellten Aufgaben Computer zuzulassen und zu verwenden. Hinzu kommt, dass einige Studien darauf hinweisen, dass die Verwendung von computerbasierten Testverfahren positive motivationale Effekte mit sich bringt (Lumsden et al., 2004; Singleton, 2001).

3.4 Vor- und Nachteile netzwerk- und internetbasierten Assessments

Die in den vorangegangenen Abschnitten aufgeführten Vor- und Nachteile von CBA sind selbstverständlich auch auf Netzwerk- und Internetbasiertes Assessment (NBA und IBA) übertragbar. Beim Testen in Netzwerkumgebungen kommen jedoch noch spezifische Vorteile, aber auch mögliche Probleme hinzu, auf die im Folgenden eingegangen wird.

3.4.1 Unkontrollierbarkeit der Testsituation

Falls netzwerkbasierte Tests ohne die Anwesenheit von Testleitern durchgeführt werden, ergibt sich das Problem, dass die Testsituation nicht kontrollierbar ist; dies ist naturgemäß beim IBT der Fall. So kann es z.B. zu „multiplen Eingaben“ („multiple submissions“) kommen, was bedeutet, dass eine Person einen Test mehrfach bearbeitet, und was somit zu einer Verfälschung der Forschungsergebnisse führen kann. Ein weiteres Problem ist das Abbrechen der Testung: Probanden können bei einer Verringerung der Motivation oder bei einer als zu hoch wahrgenommenen Testschwierigkeit die Testbeantwortung jederzeit abbrechen, ohne dass der Forscher darauf Einfluss nehmen kann oder die Gründe dafür jemals erfährt. Dies kann die Interpretation der Daten maßgeblich erschweren.

Die Unkontrollierbarkeit der Testsituation, die ein Charakteristikum von IBT darstellt, macht diese Methode für einige Einsatzbereiche psychologischer

und pädagogischer Diagnostik ungeeignet. Sobald vermutet werden muss, dass Testteilnehmer eine Motivation zur Verfälschung ihrer Ergebnisse haben, kann auf die Anwesenheit eines Testleiters nicht verzichtet werden. Für Tests, deren Ergebnisse Konsequenzen für die getesteten Personen haben (z.B. Prüfungen oder Bewerbungssituationen), scheidet IBA als Methode aus.

3.4.2 Datensicherheit

Eine für Netwerkbasiertes Testen spezifische Problematik, die sich für IBT noch schwerwiegender darstellt als für das Testen in lokalen Netzwerken, ist die Datensicherheit. Aus Datenschutzgründen ist es unbedingt notwendig, dass personenbezogene Daten, d.h. auch Testdaten, vor unbefugten Zugriffen geschützt sind. Die Frage nach der Sicherheit stellt sich ebenfalls in Bezug auf die verwendeten Testmaterialien. Sowohl aus urheberrechtlichen Gründen als auch um das Üben bestimmter Aufgaben zu verhindern, muss eine Verbreitung von Test- und Aufgabeninhalten nach Möglichkeit verhindert werden.

Zum Schutz personenbezogener Daten können die bei anderen Netzwerk-Anwendungen üblichen Techniken (Passwortschutz, Verschlüsselungsalgorithmen) eingesetzt werden. Die Sicherheit der Iteminhalte ist hingegen eine testspezifische Problematik, da die getestete Person die Aufgaben auf jeden Fall dargeboten bekommt. Es ist anzumerken, dass sich die Frage der Aufgabensicherheit grundsätzlich auch bei Papier-Bleistift-Tests stellt, aber durch die Möglichkeit einer elektronischen Vervielfältigung natürlich an Bedeutung gewinnt. Eine mögliche Vervielfältigung der Testinhalte, und sei es durch das abfotografieren von Bildschirminhalten, kann 100%ig nur durch den Einsatz von Testleitern vermieden werden, nicht jedoch durch rein technische Maßnahmen.

3.4.3 Stichprobenrekrutierung

Ein in der psychologischen Forschung häufig im Zusammenhang mit IBT diskutierter Punkt ist die Selbstselektion der Probanden in Internetuntersuchungen und die daraus resultierende, möglicherweise verzerrte Stichprobenauswahl. Tatsächlich ist es unwahrscheinlich, dass die Teilnehmer einer Internetstudie repräsentativ für eine bestimmte Zielpopulation, z.B. Deutsche einer bestimmten Altersgruppe, sind. Hierzu ist anzumerken, dass die Frage der Repräsentativität in psychologischen Untersuchungen, in denen zum großen Teil ausschließlich Psychologiestudierende untersucht werden, weitaus seltener problematisiert wird als im Zusammenhang mit per Internet rekrutierten Stichproben. Die Frage nach der Art der per Internet erreichbaren Untersuchungsteilnehmer ist tatsächlich dann bedeutsam, wenn explizit auf eine spezifische Zielpopulation generalisiert werden soll, wie dies z.B. bei Wahlprognosen der Fall ist. In den meisten psychologischen Forschungsbereichen ist die Frage der Repräsentativität weitaus weniger relevant. Per Internet rekrutierte Stichproben sind vielmehr oft heterogener als die häufig untersuchten Studierendenstichproben, womit die Generalisierung von Befunden durch IBT sogar noch gesteigert werden kann. Für die psychologische Grundlagenforschung stellt die Möglichkeit große und heterogene Stichproben zu erreichen, tatsächlich einen Hauptvorteil internetbasierter Untersuchungen dar. Die Kosten für die Testung einer großen Stichpro-

be werden reduziert, Zeitbeschränkungen und Organisationsprobleme werden verringert (Reips, 2002).

3.4.4 Technische Anforderungen

Ein problematischer Punkt, insbesondere für IBT, betrifft die Abhängigkeit von technologischen Standards. Für eine Vorgabe über das Internet müssten Testverfahren so programmiert werden, dass sie auch bei geringen Datenübertragungsraten (z.B. mit einem 56K-Modem) vorgegeben werden können. Dies bedeutet, dass die Menge der Daten so zu verringern ist, dass ein Download der nötigen Testdaten in einer dem Probanden zumutbaren Zeit stattfinden kann. Sollte dies nicht der Fall sein, könnte das wiederum zu einer Einschränkung der Stichprobe führen, da Personen, die aus verschiedenen Gründen nicht den neuesten technologischen Entwicklungen folgen, nicht mehr an einer Studie oder einem Testverfahren teilnehmen (können). Andererseits hat diese erzwungene Einschränkung der Datenmengen unter Umständen wiederum den Nachteil, dass einige der neuen Itemformate möglicherweise schlichtweg nicht verwendet werden können.

3.4.5 Gefahr mangelnder Seriosität

Ein eher nebensächlicher Punkt, der nicht die Technik des IBT selbst betrifft, hat dennoch zur Kritik an dieser Art des Testens beigetragen: Viele der so genannten psychologischen Tests, die im Internet zu finden sind, lassen berechtigte Zweifel an ihrer Ernsthaftigkeit und Glaubwürdigkeit aufkommen. Häufig sind keine Informationen über Testgütekriterien, psychometrische Merkmale, und zuweilen nicht einmal die Namen der Testautoren erhältlich. Für den Nutzer ist es daher oft schwierig „to distinguish between legitimate, professional measurement instruments and ‚pop-culture' personality quizzes, the latter being very common on the Internet" (Barak & English, 2002, S. 76).

3.4.6 Höhere zeitliche und räumliche Flexibilität der Testvorgabe

Ein offensichtlicher Vorteil des internetbasierten Assessments ist der einfache Zugang zu Testverfahren von fast überall auf der Welt. Seitens des Probanden werden dazu lediglich ein Computer sowie ein für den jeweiligen Zweck ausreichender Internetzugang benötigt. Durch computerbasiertes Assessment könnte unter Umständen ferner so genanntes „on-demand-testing" ermöglicht werden. Das würde bedeuten, dass ein Proband eine Prüfung oder einen Test dann ablegen kann, wenn er oder sie sich dazu bereit fühlt, d.h. zu dem für sie oder ihn angenehmsten und optimalen Zeitpunkt.

4 Neue Chancen bei der technologiebasierten Erfassung von Kompetenzen

Nina Jude & Joachim Wirth

Der Nutzung technologiebasierter Erhebungsverfahren in der psychologischen und pädagogischen Kompetenzerfassung kommt seit den 1990er Jahren eine immer stärkere Bedeutung zu. Parallel zum technischen Fortschritt, der Computer leicht und kostengünstig verfügbar machte und ihre Handhabung stetig optimierte, entstanden entsprechende Anwendungskontexte auch in diagnostischen Settings, die zuvor von Papier-Bleistift-Tests oder Beobachtungsverfahren geprägt waren. Während anfangs lediglich bereits existierende Verfahren, zumeist Fragebögen, direkt in eine elektronische Form umgesetzt wurden, entwickeln sich seitdem Testmodalitäten, die ohne das Medium Computer nicht denkbar wären. Folgende Vorteile der technologiebasierten Diagnostik liegen dabei auf der Hand: Eine Ökonomisierung der Erhebung und Auswertung bei gleichzeitiger Verwendung neuer Testverfahren und Aufgabenformate sowie die Möglichkeit individuell zugeschnittener Testadministration (adaptives Testen). Darüber hinaus eröffnet technologiebasiertes Testen den Zugang zu Kompetenzbereichen, die mit traditionellen Tests entweder gar nicht oder nur schwer zu erheben sind, denn es bietet die Möglichkeit der dynamischen Erfassung von Prozessabläufen, der Schaffung virtueller Testumgebungen und der Simulation lebensnaher Situationen.

Computerbasierte Testverfahren bieten die Möglichkeit einer ökonomischen Erhebung und Verwaltung von Daten. Besonders vernetzte Computer in lokalen Netzwerken oder verbunden über das Internet ermöglichen eine zeitpunkt- und standortunabhängige Vorgabe von Tests und die Rückmeldung der Testergebnisse an praktisch unbegrenzt viele Personen (ETS, 2005; Fleischer, Pallack, Wirth & Leutner, 2005; Groot, de Sonneville & Stins, 2004). Dieser Vorteil kommt nicht nur bei der individuellen Diagnostik, bspw. bei Hochschuleingangstests zum Tragen, sondern vor allem in der Anwendung bei großen Stichproben in *large scale assessments* (Hamilton, Klein & Lorie, 2000). Aus wissenschaftlicher Perspektive bieten internetbasierte Test- und Itemdatenbanken die Möglichkeit der gemeinsamen Nutzung und Entwicklung von Testitems bei gleichzeitiger Validierung und Aktualisierung der jeweiligen Testsets (Plichart, Jadoul, Vandenabeele & Latour, 2004). Neben diesen eher technischen Vorteilen eröffnen technologiebasierte Testverfahren neue Möglichkeiten für diejenigen Kompetenzbereiche, die über rein kognitive Wissensinhalte hinausgehen. Kompetenzkonzepte, in denen die Dynamik der Situation von Bedeutung ist, sind mit traditionellen Verfahren nur schwer zugänglich. Meistens werden Verhaltensbeobachtungen aufwändig codiert und analysiert. Durch die Integration von multimedialen Testinhalten wie Film- und Videosequenzen oder auch interaktiven Simulationen können authentische und lebensnahe Testumgebungen gestaltet werden (Drasgow, 2002). Dies bietet den Vorteil höherer Validität im

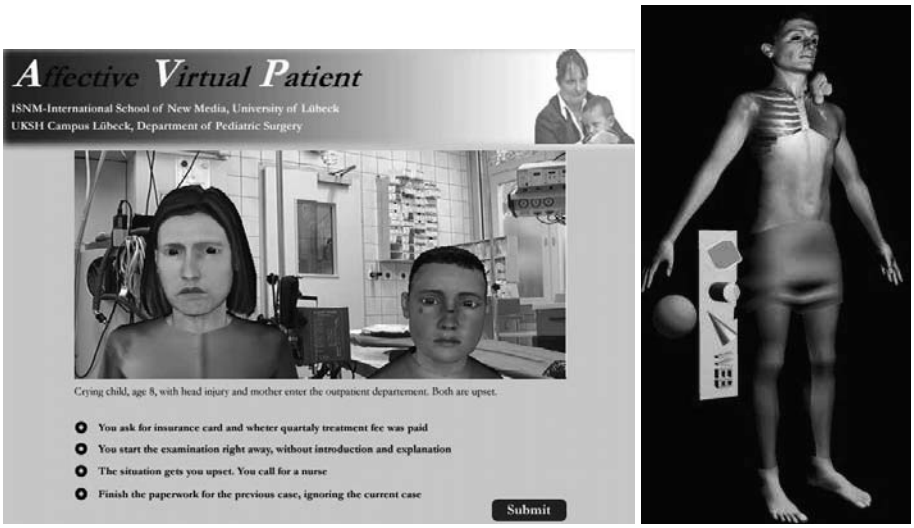
Bezug auf den zu erhebenden Kompetenzbereich, als dies bei traditionellen Testverfahren der Fall wäre. Durch den Einbezug adaptiver Testroutinen können zudem beliebig offene, prozessorientierte Testsituationen gestaltet werden, die dem Handeln in realen Situationen sehr nahe kommen. Im Folgenden werden diese Chancen der technologiebasierten Erfassung von Kompetenzen an verschiedenen Anwendungsbeispielen expliziert.

4.1 Erfassung komplexer und dynamischer Kompetenzen

Ein Kompetenzbereich, der sich aufgrund seiner Mehrdimensionalität und Dynamik besonders für den Einsatz computerbasierter Testverfahren eignet, ist die komplexe Problemlösefähigkeit (Dörner, Kreuzig, Reither & Stäudel, 1983; Dörner & Preußler, 1990; Dörner, Schaub & Strohschneider, 1999). Bekannt wurde die deutsche Forschergruppe um Dörner mit ihren Arbeiten zu der hochkomplexen und dynamischen Computersimulation von "Lohhausen", einer virtuellen Stadt, die von den zu testenden Personen verwaltet werden musste. Die Testperson wurde also zum Bürgermeister in einer kleinen virtuellen Stadt, deren Simulation auf ca. 2000 miteinander vernetzten Parametern basierte. Jeder Eingriff des Bürgermeisters wirkte sich direkt und indirekt auf eine oder mehrere dieser Parameter und damit auf die Entwicklung der virtuellen Stadt aus. Das Erkennen der komplexen Zusammenhänge und die positive Beeinflussung der Stadtentwicklung waren die Aufgaben, die es zu bearbeiten galt. Die Erkenntnisse dieses Testsystems hatten und haben einen wegweisenden Einfluss auf die Erforschung von Problemlösekompetenz, so dass computerbasierte Tests aus diesem Forschungsbereich nicht mehr wegzudenken sind (Baker & O'Neil, 2002; Klieme, Leutner & Wirth, 2005; Wirth & Klieme, 2003).

Weitere spezifische Anwendungsfelder computerbasierter Diagnoseverfahren finden sich besonders in den Bereichen beruflich relevanter Kompetenzen. Während berufliches Fachwissen ggf. noch durch Papier-Bleistift-Verfahren abzufragen ist, kommt dem beruflich kompetenten Handeln in dynamischen, komplexen und interaktiven Situationen eine große Bedeutung zu. Um dieses zu erfassen und zu bewerten sind aufwändige Beobachtungsverfahren die klassische Methode der Wahl, bei der Rateeffekte jedoch nicht auszuschließen sind. Zudem ist es schwierig, solche Situationen standardisiert jeder zu testenden Person vorzugeben. Für die computerbasierte Erfassung beruflicher Handlungskompetenz bieten sich daher dynamische und interaktive Programme an, die berufsspezifische, standardisierte Situationen mit verschiedenen Handlungsoptionen simulieren (Strauß & Kleinmann, 1995; Streufert, Pogash & Piasecki, 1988; Wagener, 2001). Häufige Anwendung finden solche Verfahren besonders in den medizinischen Ausbildungsfeldern (Clyman, Melnick & Clauser, 1995; vgl. Luecht & Clauser, 2002), wo sie sowohl zur Überprüfung der Kompetenz als auch zu Übungszwecken eingesetzt werden (vgl. Abbildung 4.1).

Abbildung 4.1: Testaufgabe der Affective Virtual Patient – Simulation von Jung, Ahad & Weber (2005) und der Dynamic Virtual Patient – Simulation von Gauthier (2005).



Die Abbildung zeigt Ausschnitte aus zwei verschiedenen medizinischen Simulationen. Der Affective Virtual Patient (linkes Bild; Jung, Ahad & Weber, 2005) stellt den angehenden Mediziner vor unterschiedlich komplexe Aufgaben, wie die Neuaufnahme und Untersuchung eines unbekannten Patienten, die in Teilschritten anzugehen sind. Je nach Entscheidung, die über eine multiple-choice Abfrage erfasst wird, ändert sich die folgende Situation, passt sich also dynamisch an. Werden in einem Schritt fehlerhafte Entscheidungen getroffen, wirken sich diese auf das nachfolgende Geschehen aus. Diese Aufgaben erfragen nicht nur medizinisches Wissen, sondern überprüfen facettenreiches Handeln in komplexen Situationen, die bspw. auch die soziale Kompetenz der zu testenden Person und das Wissen um Routinen im Krankenhausalltag einbeziehen. Die Simulation von Gauthier (2005) fokussiert stärker auf medizinische Aspekte, in dem Symptome und (chirurgische) Behandlungen spezieller Erkrankungen im Mittelpunkt der Kompetenzerfassung stehen. Beide Verfahren zeigen die besonderen Vorteile der technologiebasierten Kompetenzerfassung: Die Testsituation ist für jeden Probanden vergleichbar, es werden keine echten Patienten für die Teilnahme benötigt, vielmehr können virtuelle Patienten mit bestimmten Eigenschaften erschaffen werden, die beliebig oft einsetzbar sind. Auch spezielle Krankheitsbilder, die in der Realität ggf. selten auftreten, können problemlos dargestellt werden. Diese Testverfahren können auch unerwartete Notfälle simulieren und die folgenden Reaktion des Arztes sowie die Angemessenheit der virtuellen Behandlung erfassen. Alternativ zu reinen Computersimulationen werden Videovignetten von arbeitsplatzspezifischen Situationen computergestützt vorgegeben, das Antwortverhalten der Testpersonen kann dann bezogen auf unterschiedliche Kriterien – pragmatische Angemessenheit, Korrektheit der Antworten, Nutzung von Fachvokabular – ausgewertet und rückgemeldet werden (Peabody et al., 2004; Van den Branden, Depauw & Gysen,

2002). Solche Simulationen als Testszenarien berücksichtigen die Situationsgebundenheit von Kompetenzen und erhöhen somit die Testvalidität, wodurch bessere Rückschlüsse von der Leistung im Test auf tatsächliche Kompetenz möglich werden.

Doch auch klassische Kompetenzbereiche können mittels Computertechnologie effizienter, valider und messgenauer erfasst werden, wie bspw. die verbale produktive Sprachkompetenz, die in traditionellen Testverfahren aus Kostengründen meist vernachlässigt wurde. Ein Beispiel hierfür stellen die aktuellen Verfahren des *Test of English as a Foreign Language* (TOEFL) dar (ETS, 2005). Die neue, internetbasierte Testversion TOEFL-iBT nutzt digitale Aufnahmetechniken und deren internetbasierte Vermittlung, um Daten einer großen Stichprobe schnell und objektiv zu erfassen und zu bewerten. Auch für spezielle Zielgruppen, bei denen der Umgang mit schriftbasierten Fragestellungen oder Testheften eher abschreckend wirken könnte und deren Kompetenz in klassischen Testsituationen daher oft nicht valide zu erfassen ist, eröffnen computerbasierte Testverfahren neue Möglichkeiten. Hierzu gehören sicherlich Kinder, die durch einfache, ansprechend gestaltete Testvorgaben am Computer zur Teilnahme motiviert werden können (vgl. Konak, Duindam & Kamphuis, 2005).

Abbildung 4.2: Testbildschirm aus dem CITO-Test zur Erfassung der Sprachkompetenz im Vorschulalter (CITO, 2005).

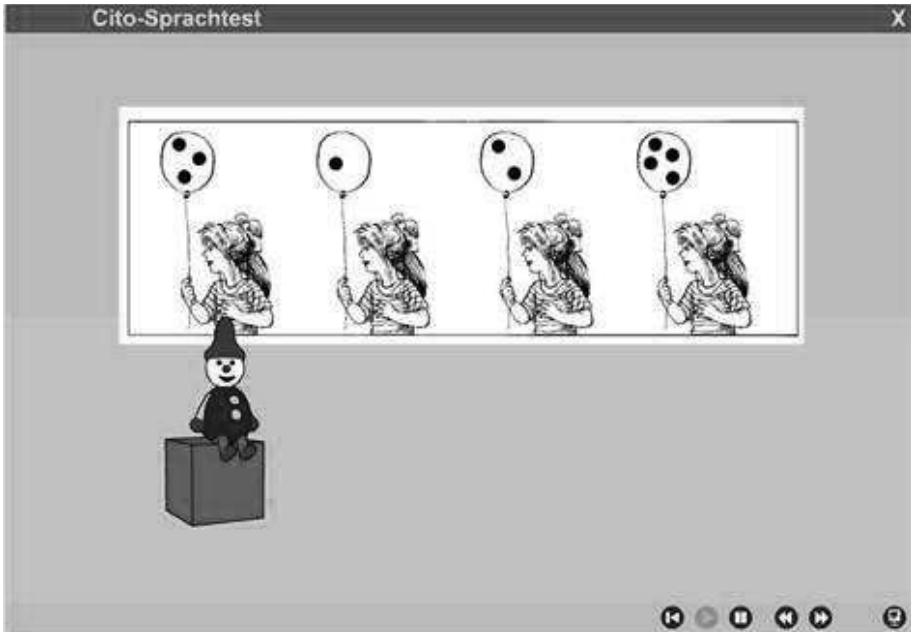


Abbildung 4.2 zeigt einen Bildschirmausschnitt aus dem CITO Sprachtest (CITO, 2005) zur Erfassung der Sprachkompetenz im Vorschulalter, der von Kindern unter Aufsicht eines Testleiters selbständig am Computer bearbeitet werden kann. Kindgerecht gestaltete sprachliche Situationen werden von einem Clown kommentiert und mit auditiven Aufforderungen versehen, bspw. „*Klicke mit der Computermaus auf den Ballon mit drei Punkten*". Dieses Verfahren erweist sich in

der Praxis als besonders objektiv, da die Angemessenheit von Antworten keinen Testleitereffekten unterliegt. Der Anreiz zur Teilnahme auf Seiten der Kinder ist hoch, denn die Gestaltung der Testsituation ähnelt einem Computerspiel. Zudem ist eine schnelle Testauswertung und Rückmeldung des Sprachstands problemlos möglich.

4.2 Multimediale Aufgaben- und Testformate

Die größte Neuerung zwischen traditionellen und computerbasierten Testverfahren stellt die unkomplizierte Verwendung von multimedialen und dynamischen Teststimuli wie Audiodateien, Videos oder Animationen dar, die inzwischen in vielfältigen Bereichen der Kompetenzerfassung zu finden sind. Dazu gehören u.a. die Bereiche der Musik, Geschichte und spezielle der Lehrerausbildung (e.g., Ackermann, Evans, Park, Tamassia & Turner, 1999; Bennett et al., 1999; Krauss et al., 2004; Olson-Buchanan et al., 1998; Vispoel, 1999). Die Stimuli sind frei von Umgebungseinflüssen und können individualisiert vorgegeben werden, je nach Testaufgabe können bspw. Videosequenzen individuell gestoppt oder wiederholt werden und Audiodateien für jeden Testtaker individuell eingespielt werden (Bennett et al., 1999; Krauss et al., 2004). Dadurch können auch Unterschiede in der individuellen Informationsverarbeitung durch die Testsituation aufgefangen werden, um Testfairness zu gewährleisten. Multimediale und dynamische Stimuli bieten darüber hinaus eine realitätsnahe und authentischere Gestaltung von Testsituationen. Dazu gehören auch Verfahren, die nicht beobachtbare Prozesse mittels Animationen visualisieren – bspw. den Ablauf von Stoffwechselprozessen (Nerdel, 2003). Auch Experimente, die in realen Umgebungen aus Sicherheitsgründen nicht durchgeführt werden könnten, lassen sich am Computer leicht simulieren (Mikelskis, 1997; Prenzel, von Davier, Bleschke, Senkbeil & Urhahne, 2000). Der Gestaltung einer solchen multimedialen Testumgebung sind dabei kaum Grenzen gesetzt.

Aktuelle Studien beschäftigen sich mit der Gestaltung dreidimensionaler Testumgebungen. In diesem kann sich die zu testende Person frei bewegen, Handlungen ausführen oder auch Gespräche mit simulierten bzw. echten Personen führen. Die Handhabung solcher komplexer, virtueller Testumgebungen ist intuitiv bzw. schnell erlernbar (Frey, Hartig, Ketzel, Zinkernagel & Moosbrugger, im Druck). Die sehr realitätsnahen Szenarien lassen sich dabei für psychologische Experimente, zur Einzeldiagnose von Kompetenz und auch zur Beobachtung und Aufzeichnung von sozialen Interaktionen nutzen. Der Gestaltung solcher virtueller Umgebungen sind dabei wenig Grenzen gesetzt: Eine virtuelle Fabrik, in der verschiedene Fertigungsstationen aufgesucht werden können, ist eben so zu gestalten wie eine Gruppendiskussion, in der reale und virtuelle Personen ihre Meinung zu einem bestimmten Thema vertreten müssen. Besonders bei der Erfassung von sozialen Kompetenzen bieten solche technologiebasierten Verfahren die reizvolle Möglichkeit, verbales und nonverbales Verhalten direkt aufzuzeichnen, ohne bspw. die Codierung von Videomaterial oder das Transkribieren von Gesprächsprotokollen. Interaktionsabläufe können so schnell und effizient gespeichert und analysiert werden. Ein Beispiel

hierfür findet sich in der Simulation *Psi-Land* (Frey, Blunk & Banse, 2006, vgl. Abbildung 4.3).

Abbildung 4.3: Bildschirmausschnitt der Simulation Psi-Land (Frey, Blunk & Banse, 2006).

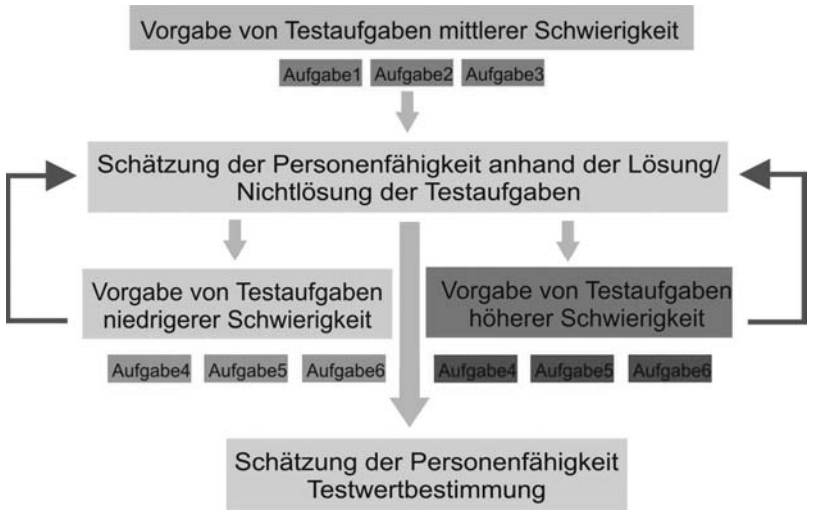
In der virtuellen Umgebung von *Psi-Land* können Personen sich mittels virtueller Darstellung ihrer selbst bewegen, unterhalten und gemeinsam oder gegeneinander Testaufgaben lösen. Nicht nur die Bewegungsabläufe, sondern auch die Anzahl der verbalen Interaktionen, der Blickkontakt der virtuellen Figuren werden aufgezeichnet und z.B. mit externen Maßen wie Persönlichkeit oder gegenseitigen Einstellungen in Zusammenhang gebracht. Dieses noch in den Anfängen seiner Entwicklung befindliche Verfahren, realistische Testszenarien für mehrere Personen zu schaffen, ließe sich für eine vernetzte Nutzung weiterentwickeln, so dass Testteilnehmer nicht an Computern vor Ort, sondern durch das Internet verbunden gemeinsam an Aufgabenstellungen arbeiten können. Solchermaßen multimediale Testszenarien bieten dann die Möglichkeit, kognitive Kompetenzen, Verhalten und interpersonale Kommunikation gleichermaßen zu erfassen.

4.3 Datenerfassung, Verwaltung und Rückmeldung

Die meisten traditionellen Testverfahren verwenden Papier-Bleistift-Tests, in denen die zu testenden Personen Antworten schriftlich niederlegen oder im einfachsten Fall nur ankreuzen. Allein durch diesen Vorgabemodus ist der gezeigte Verhaltensbereich maximal eingeschränkt und realitätsfern, potenzielle Störeinflüsse wie allgemeine Schreibfähigkeit müssen berücksichtigt werden. Im Gegensatz dazu bieten computerbasierte Verfahren eine Vielzahl von Aufzeichnungsgeräten, die Verhaltensdaten in Echtzeit protokollieren. Neben den

klassischen Komponenten wie Tastatur und Maus sind Bestandteile von Fahr- oder Flugzeugen wie Pedale, Lenkräder oder echte Schaltkonsolen problemlos in computergesteuerte, elektronische Testumgebungen zu integrieren (Harris & Khan, 2003). Blickrichtungen und Dauer können zur Erfassung der Aufmerksamkeit über Kameras ebenso registriert werden wie verbale Äußerungen über Mikrofone (ETS, 2005; Norman, Debus, Doerre & Leutner, 2004; Rosenblum, Parush & Weiss 2003a, 2003b). Die Eingabegeräte können dem jeweiligen Kompetenzbereich angepasst werden und auch Personen, die im Umgang mit Tastatur oder Maus unerfahren sind, werden nicht benachteiligt, wenn die Erfassungsgeräte ihren Fähigkeiten angepasst werden: So ist es einfacher, mit dem Finger auf einen berührungssensiblen Touchscreen zu drücken, als mit der Computermaus einen bestimmten Punkt auf dem Bildschirm anzupeilen. Besonders auf technikferne und ältere Zielgruppen ist hierdurch Rücksicht zu nehmen. Unabhängig von der Art und Anzahl der Eingabegeräte ist die Datenerfassung der Einzelgeräte über den Computer gekoppelt, der nicht nur die eingehenden Daten speichert, sondern im Idealfall simultan auswertet, um z.B. die Testszenarien zu modifizieren oder neue Aufgaben vorzugeben, ohne den Testprozess als solchen zu unterbrechen (Hadwin & Winne, 2001; Winne, 1982; Winne, Jamieson-Noel & Muis, 2002). Ein Vorteil dieser gleichzeitigen Datenerhebung und -analyse besteht in der Möglichkeit des sog. *adaptiven Testens*: Die Vorgabe von Testaufgaben kann so gesteuert werden, dass diese dem individuellen Kompetenzniveau der zu testenden Person angepasst werden (vgl. Abbildung 4.4).

Abbildung 4.4: Schematische Darstellung eines adaptiven Testvorgangs.



Ausgehend von Aufgaben unterschiedlicher, mittlerer Schwierigkeiten wird anhand des Lösungsmusters die Personenfähigkeit geschätzt. Sind die Aufgaben zu schwer, wurden also nur wenige gelöst, werden leichtere Aufgaben gewählt, die in einem niedrigeren Kompetenzbereich besser differenzieren – sind die Aufgaben zu leicht, werden entsprechend schwerere gewählt. So können jeweils diejenigen Aufgaben ausgewählt werden, die am meisten Information

über die Fähigkeit liefern (Lord, 1980). Adaptive Tests sind psychometrisch somit effizienter und entsprechend weniger zeitaufwändig (Folk & Smith, 2002). Die Schätzung der Personenfähigkeit und die Auswahl der nächsten Testaufgaben erfolgt in Sekundenbruchteilen, gesteuert von Algorithmen, die während der Testdurchführung simultan im Hintergrund ablaufen. Sobald der jeweilige Testdurchlauf beendet ist, stehen die Ergebnisse direkt zur Auswertung zur Verfügung und können bei Bedarf auch der getesteten Person sofort zurückgemeldet werden. Hierbei können nicht nur Gesamttestwerte verwendet werden, auch einzelne Kompetenzbereiche bis hin zu Antworten auf einzelne Testitems sind bei Bedarf unmittelbar einsehbar. Nützlich ist dies besonders in Beratungssituationen, beispielsweise bei der Berufsberatung oder der Hochschulzulassung.

4.4 Messgütekriterien

Um technologiebasierten Erhebungsverfahren einen Vorteil gegenüber klassischen Tests zuschreiben zu können, müssen sie die wesentlichen Messgütekriterien mindestens genauso gut, im optimalen Fall noch besser erfüllen. Das Gütekriterium der *Objektivität*, also der Unabhängigkeit der Messung und des Testwertes von Art und Ort der Durchführung, der Art der Auswertung sowie der Interpretation, ist dabei am leichtesten zu erreichen. Je mehr Abläufe durch den Computer kontrolliert werden, umso weniger Einfluss durch Testleitung oder Testumgebung sind anzunehmen. Auch die Auswertung und Interpretation der Testwerte läuft automatisch und differenziert dokumentiert ab, sie ist somit jederzeit nachvollziehbar. Sogar offene Antworten oder kurze freie verbale wie nonverbale Texte können inzwischen automatisch ausgewertet werden, so dass zufällige Messfehler oder die Gefahr subjektiver Ergebnisinterpretation reduziert werden. Entsprechende Testwerte übertreffen in ihrer Objektivität solche, die durch traditionelle und subjektive Ratingverfahren gewonnen werden (Breland & Lytle, 1990; Burstein, Kaplan, Wolff & Lu, 1997; Chung, Baker, Brill, Sinha & Saadat, 2003; Chung, O'Neil, Bewley & Baker, in diesem Buch; Franzke, Kintsch & Kintsch, 2005; Page & Petersen, 1995). Zusätzlich ist diese Art der Auswertung schneller und kostengünstiger als der Einsatz von menschlichen Auswertern. Die *Reliabilität* computerbasierter Verfahren unterscheidet sich im Normalfall nicht von der vergleichbarer Papier-Bleistift-Tests (Herl, O'Neil, Chung & Schacter, 1999; Pinsoneault, 1996). Insbesondere der Transfer von Persönlichkeits-, Einstellungs- und Interessentests in computer- oder internetbasierte Verfahren beeinträchtigt weder die Zuverlässigkeit der Messung noch die inhaltliche Validität (Ridder, Bruns & Brünn, 2004). Bezüglich breiter Kompetenzbereiche, die besonders Verhaltensaspekte in den Vordergrund stellen, sind bei der Überprüfung der *Validität* von technologiebasierten Verfahren entsprechend umfassende Zugänge zu wählen. So ist nicht nur der Vergleich zu traditionellen Tests zu ziehen sondern auch zu analysieren, inwieweit beispielsweise simulierte Situationen mit den realen übereinstimmen, um zu recht vom Testverhalten auf die tatsächliche Kompetenz schließen zu können. Letztendlich gilt für klassische ebenso wie für technologiebasierte Testverfahren der Anspruch, dass Messgütekriterien kritisch zu überprüfen sind. Vor diesem Hintergrund eröffnen neue, computergestützte Methoden viel versprechende Möglichkeiten für die Erfassung von Kompetenzen.

5 Anforderungen an computer- und netzwerkbasiertes Assessment

Johannes Hartig, Ulf Kröhne & Astrid Jurecka

5.1 Computer- und internetbasierte Diagnostik in der Bildungsforschung

Computertechnologie ist in den letzten Jahrzehnten zu einem unverzichtbaren Werkzeug pädagogischer und psychologischer Diagnostik geworden. Zunächst wurden vor allem Computerversionen von Tests erstellt, die ursprünglich als Papier-Bleistift-Versionen entwickelt worden waren. Mittlerweile findet das neue Medium bereits breite Anwendung für diagnostische Techniken, die ohne Computereinsatz nicht realisierbar wären. Beispiele sind das Computerbasierte Adaptive Testen (CAT) oder der Einsatz von interaktivem Aufgabenmaterial.

Zu den neuen diagnostischen Techniken, die sich durch den Einsatz von Computern ergeben, kommen in den letzten Jahren weitere Möglichkeiten durch den Einsatz von diagnostischen Verfahren auf vernetzten Computern hinzu. Besonders attraktiv ist hierbei die Verwendung des Internet bei der Anwendung von computerbasierten Tests. Die Testdurchführung wird räumlich und zeitlich von einem Testleiter unabhängig, große Teilnehmerkreise können mit geringem ökonomischem Aufwand erreicht werden. Insbesondere bei psychologischen Fragebogenuntersuchungen in der Grundlagenforschung ist die Datenerhebung im Internet zu einer etablierten Methode geworden, um große und heterogene Stichproben zu rekrutieren. Jedoch auch im Bereich der Leistungsdiagnostik gibt es bereits vielfältige Ansätze, sich das Internet zur Darbietung von Tests und zur Sammlung diagnostischer Daten zunutze zu machen. So kommen z.B. in mehreren US-Bundesstaaten internetbasierte Tests im Bildungsbereich zum Einsatz: 18 US-Staaten wenden zumindest zum Teil computerbasiertes Assessment im Bildungsbereich an, weitere führen zurzeit Pilotprojekte durch. Der Staat mit dem vermutlich umfassendsten Online-Assessment-Programm ist Virginia. Dort werden fast alle standardisierten Tests in der High School auf diesem Wege durchgeführt (Bennett, 2005). Auch in einigen europäischen Staaten wie beispielsweise Schottland und England wurden bereits in den letzten Jahren weit reichende Initiativen für die Entwicklung und Erforschung computerbasierter Testsysteme im Bildungsbereich gestartet.

In den meisten der oben genannten Länder sind landes- oder bundesweit standardisierte Test- und Prüfverfahren im Bildungswesen seit langer Zeit Tradition. Dort ist die größte Motivation für den Einsatz von computerbasiertem Assessment meist die Übertragung der bisher auf Papier-Bleistift-Basis vorgegebenen Testverfahren in ein elektronisches System aus zeitökonomischen und finanziellen Gründen.

In Deutschland hingegen war die Durchführung zentral standardisierter Testverfahren im Bildungsbereich lange Zeit unüblich. Es zeichnet sich jedoch ab, dass sie auch im deutschen Bildungssystem in Zukunft eine zunehmend größere Rolle spielen. Zu internationalen und nationalen Schulleistungsstudien kommt die kontinuierliche Überprüfung der Bildungsstandards hinzu. In allen deutschen Bundesländern wurden Qualitätsentwicklungs-Institute eingerichtet, ebenso ein zentrales Institut für Qualitätsentwicklung im Bildungswesen (IQB) durch die Kultusministerkonferenz. Im Zuge dieser zunehmenden Bedeutung standardisierter Tests dürfte auch die Anwendung computer- und netzwerkbasierter Diagnostik in Deutschland in Zukunft eine wichtigere Rolle spielen.

In Deutschland kommen computerbasierte Testanwendungen derzeit vor allem in der psychologischen Forschung und teilweise in der psychologischen Berufseignungsdiagnostik, im Bildungswesen jedoch kaum in größerem Maßstab zum Einsatz. Die Entwicklungsbedingungen dieser Anwendungen, soweit sie nicht ganz von ausländischen Anbietern eingekauft werden, sind sehr unterschiedlich und zum großen Teil suboptimal. Häufig werden diagnostische Instrumente in Forschungseinrichtungen entwickelt, die von ihrer inhaltlichen Ausrichtung und Ausstattung nicht über die Kompetenzen und Kapazitäten für eine professionelle Softwareentwicklung verfügen – dies gilt teilweise sogar für Software, die später von deutschsprachigen Testverlagen kommerziell vertrieben wird. Die Softwareentwicklung wird entweder von kurzzeitig angestelltem Personal wie z.B. Werksstudenten vorgenommen, oder anderweitig qualifiziertes wissenschaftliches Personal arbeitet sich kurzfristig in einen bestimmten technischen Bereich ein. Die unter diesen Bedingungen entwickelten Produkte entsprechen häufig nicht dem aktuellen technischen Stand; eine angemessene technische Dokumentation ist häufig nicht gewährleistet. Die Produkte, die mit begrenzten Ressourcen und in der Regel für eine konkret anstehende spezielle diagnostische Fragestellung entwickelt werden, sind häufig für andere Fragestellungen nicht zu verwenden. Auch sind verschiedene Entwicklungen untereinander meist nicht kompatibel bzw. ineinander überführbar. Als Folge dieser Entstehungsbedingung werden sie oft nicht gepflegt und weiterentwickelt. Angesichts der raschen technischen Entwicklungen im Hard- und Softwarebereich ist die so entwickelte diagnostische Software innerhalb relativ kurzer Zeit veraltet und z.B. auf Rechnern mit neueren Betriebssystemen nicht mehr einsetzbar. Eine Folge der hier skizzierten Entwicklungsbedingungen diagnostischer Software ist, dass häufig Verfahren zum Einsatz kommen, die verglichen mit professioneller Software deutlich veraltet sind und hinter den aktuellen technischen Möglichkeiten weit zurück bleiben.

Hinzu kommt, dass für neue Fragestellungen in der Regel neue Software entwickelt werden muss, da auf bereits vorliegende Produkte kaum aufgebaut werden kann. Dies bedeutet, dass der Produktionsprozess, der ohnehin oft unter schwierigen personellen Voraussetzungen stattfindet, immer wieder bei Null anfängt. Es ist zu vermuten, dass grundlegende Entwicklungsschritte, die vielen Anwendungen gemeinsam sind, hierdurch in verschiedenen Kontexten und an verschiedenen Forschungseinrichtungen immer wieder von vorne durchlaufen werden.

5.2 Chancen eines Basissystems für computer- und netzwerkbasierte Diagnostik

Angesichts der voraussichtlich zunehmenden Rolle computer- und netzwerkbasierten Assessments erscheint es angezeigt, Überlegungen zur Verbesserung der bisher suboptimalen Entwicklungsbedingungen für entsprechende Anwendungen im deutschsprachigen Raum anzustellen. Eine derartige Überlegung, mit der sich dieses und das folgende Kapitel intensiver befassen sollen, ist die Entwicklung eines *Basissystems für computer- und netzwerkbasiertes Assessment*. Die Idee ist hierbei die Entwicklung einer grundlegenden technischen Basis, auf der konkrete Instrumente für spezifische Fragestellungen aufbauen können, ohne dass die grundlegendsten Entwicklungsschritte jedes Mal neu durchlaufen werden müssen. Ein derartiges System sollte flexibel verwendbare Grundstrukturen für Datenverwaltung, Speicherung und Auswertung in diagnostischen Anwendungen bieten – der Einsatz wäre für Fragebogenuntersuchungen ebenso denkbar wie für Leistungstests. Es sollte die technischen Grundlagen für die datenbankbasierte Verwaltung und Bearbeitung sowohl von Test- und Aufgabeninhalten als auch von personenbezogenen Daten einschließlich der bei einer Testung erhobenen Reaktionen beinhalten. Ebenso sollten die Grundlagen für die Implementierung verschiedener Auswertungsalgorithmen, die Auswertung individueller oder gruppenbezogener Testdaten und darauf basierender Ergebnisrückmeldungen inbegriffen sein.

Ein zentrales Motiv für ein derartiges Vorhaben ist es, den Aufwand für Softwareentwicklung an Forschungseinrichtungen, die dafür nicht ausgestattet sind, zu reduzieren. Damit können vorhandene personelle Ressourcen stärker in die eigentliche inhaltliche Entwicklungsarbeit investiert werden. Die Verwendung eines Basissystems für möglichst viele Assessment-Aktivitäten würde ferner einen inhaltlichen Austausch von Ergebnissen, Items, Daten und Teilentwicklungen innerhalb einer Anwendergemeinde begünstigen. Alle Forscher, die mit einem derartigen System arbeiten würden, bzw. die Programmierung von beispielsweise neuen Itemformaten beauftragen, würden dieselben Standards bei der Entwicklung beachten. Dies würde die Kommunikation zwischen Forschern bzw. Programmierern stark vereinfachen.

Die Entwicklungsarbeit bei einem neuen Testverfahren würde sich auf Seiten der Forscher wieder auf inhaltliche Entwicklungen konzentrieren und fokussieren können, anstatt jedes Mal die Notwendigkeit mit sich zu bringen, sich neues technisches Hintergrundwissen und Grundlagen aneignen zu müssen. Es handelt sich hierbei also um den Wunsch nach einem grundlegenden System, das Anwendern in der psychologischen und pädagogischen Diagnostik die Entwicklung eigener Instrumente erleichtert.

Eine wichtige Voraussetzung hierfür wäre eine entsprechende öffentlich zugängliche Dokumentation der Entwicklung und Anwendung von Tests und Testteilen, wobei es den Entwicklern allerdings freigestellt bleiben sollte, inwieweit sie die für einen speziellen Test entwickelten tatsächlichen Items freigeben möchten. Ein Beispiel wäre hier ein neues Muster (*template*), das für ein

bestimmtes neues Itemformat entwickelt wurde, öffentlich zugänglich zu machen und zu dokumentieren, allerdings die Items selbst nicht der breiten Öffentlichkeit zugänglich zu machen. Auch der Umgang mit den eigenen erhobenen Daten muss dem Forscher bzw. dessen Auftraggeber selbstverständlich freigestellt sein.

In den folgenden Abschnitten wird die Idee eines Basissystems für netz- und computerbasiertes Assessment anhand der technischen und inhaltlichen Anforderungen, die an ein derartiges System zu stellen wären, illustriert. Die organisatorischen und politischen Rahmenbedingungen, die für eine derartige Entwicklung und vor allem den Aufbau und die Koordination einer hinreichend großen Nutzergemeinde nötig sind, sind nicht Gegenstand dieses Bandes.

5.3 Anforderungen an ein Basissystem für netzwerk- und computerbasiertes Assessment

Im Folgenden sollen wesentliche technische Anforderungen skizziert werden, die sich an ein flexibles Basissystem für computer- und netzwerkbasiertes Assessment formulieren lassen. Zunächst sollen allgemeine Anforderungen an die Datenverwaltungsstrukturen eines derartigen Systems skizziert werden, die für Netzwerkanwendungen generell bedeutsam sind. Neben solchen grundlegenden strukturellen Anforderungen ergeben sich weitere spezifische Anforderungen aus dem konkreten Anwendungskontext pädagogisch-psychologischer Diagnostik. Diese, von praktischer Seite entstehenden Anforderungen, werden im Einzelnen in einem weiteren Abschnitt skizziert. Auch sie können Implikationen für die Struktur der verarbeiteten Daten haben, aber auch für spezifische Features und Charakteristika einer möglichen Programmoberfläche.

5.3.1 Allgemeine strukturelle Anforderungen

Flexibilität hinsichtlich von Datenformaten und Datenschnittstellen

Die wichtigste Anforderung an ein Basissystem für computer- und netzwerkbasiertes Assessment ist Flexibilität, durch die es hinreichend offen für die Anpassungen an verschiedene Anwendungen in unterschiedlichen Kontexten ist. Ein zentraler Punkt in dieser Hinsicht betrifft den Datenaustausch mit anderen Systemen und Anwendungen. Bei der Entwicklung sollte angestrebt werden, das System zum einen direkt kompatibel mit einer möglichst großen Zahl absehbar notwendiger Datenformate zu machen, zum anderen aber auch die Möglichkeit zukünftiger Erweiterungen noch offen zu lassen.

Kompatibilität ist unter anderem wichtig im Austausch mit anderen Anwendungen, die im Zusammenhang mit einer spezifischen diagnostischen Anwendung benötigt werden. So ist es z.B. vorstellbar, dass an der direkten Schnittstelle zur getesteten Person spezielle zusätzliche Anwendungen verwendet werden, wenn die Reaktionen auf den Test über spezifische Hardware erfolgen.

Noch wichtiger ist die Flexibilität eines solchen Basissystems im Kontext netzwerkbasierten Assessments. Es sollte möglich sein, sowohl personen- als auch aufgabenbezogene Daten zwischen dem Testsystem und anderen Systemen dynamisch auszutauschen.

Schließlich sollte das System auch die Voraussetzungen für einen nutzergesteuerten Import und Export von Daten aus in der Forschung häufigen Formaten spezifischer Anwendungen bieten, z.B. gängiger Statistik- oder Tabellenkalkulationsprogramme.

Flexible Verwendung von Netzwerkstrukturen

Das hier skizzierte Basissystem sollte verschiedene Formen der Nutzung von Netzwerken zur Testdarbietung und Datenspeicherung ermöglichen, um in möglichst vielen Forschungs- und Anwendungskontexten nutzbar zu sein (vgl. auch Anwendungsszenarien in Kapitel 6). Auf einem Kontinuum von einer Testung auf einem einzelnen Rechner ohne Netzwerkverbindung bis zu einer öffentlich zugänglichen internetbasierten Testung mit einem zentralen Testserver sollte jede Art der Netzwerknutzung möglich sein. Das System könnte zu diesem Zweck so gestaltet sein, dass es sich bei einer lokalen Anwendung auf einem virtuellen Server auf dem Computer, auf welchem auch die Testdarbietung erfolgt, einrichten lässt; bei einer breiteren Netwerkanwendung auf einem dezidierten Server bis hin zu einem Cluster von Servern.

Eine wichtige Anforderung im Zusammenhang einer verteilten Testdarbietung mit verschiedenen Stufen der Netzwerknutzung ist die Koordination verschiedener Versionen derselben Datenstruktur. Auch wenn die Testdarbietung lokal auf einzelnen Computern oder in kleineren lokalen Netzwerken mit jeweils eigenen Servern (z.B. in einer Schule oder Schulklasse) erfolgt, sollte ein Zusammenführen der Daten zu einer zentralen Datenbank möglich sein. Hierbei sind nicht nur die bei der Testung erfassten Antworten zu berücksichtigen. Auch das lokale Hinzufügen und Ändern von Personendaten (z.B. von Schülern) und aufgabenbezogener Daten (z.B. zur Kontrolle der Itemexposition[4]) muss mit einer zentralen Version der Anwendung, in der die Gesamtheit der Daten für einen Test verwaltet wird, synchronisiert werden können.

Flexible Organisation von Administrator- und Teilnehmerrechten

Die Vielfalt möglicher Anwendungskontexte eines Systems für computer- und netzwerkbasiertes Assessment stellt auch Anforderungen an die Flexibilität von Administrator- und Nutzerrechten. Zwischen Administratoren, die vollen Zugriff auf Personen- und Aufgabendaten haben, und reinen Testteilnehmern, die sich lediglich zur Identifikation am System authentifizieren sind eine ganze Reihe weiterer spezifischer Nutzerprofile denkbar. So könnte es z.B. wünschenswert sein, Lehrern Zugriff auf Testergebnisse ihrer Klasse oder auch einzelner Schü-

[4] Mit Itemexposition ist gemeint, wie häufig ein einzelnes Item bei einer größeren oder länger laufenden Studie vorgegeben wird. Dies soll häufig kontrolliert werden, damit Items, die auch in zukünftigen Studien eingesetzt werden sollen, nicht bekannt werden.

ler zu geben, nicht jedoch auf Aufgabeninhalte; Schulleiter könnten Zugriff auf Ergebnisse auf Schulebene bekommen. Wenn diese Lehrer oder Schulleiter ihrerseits z.B. fragebogenbasierte Informationen in das System einspeisen, haben sie im System selbst die Rolle von „Testteilnehmern". In spezifischen Lehr- oder Forschungskontexten kann es auch sinnvoll sein, den Testteilnehmern begrenzten Zugriff auf eigene, aber auch gruppenbezogene Testergebnisse zu gewähren. In Forschungskontexten muss zudem eine Selbstanmeldung von interessierten Teilnehmern als Nutzer des Systems, d.h. das Neueinrichten von Benutzerkonten, möglich sein. Ein Basissystem für computer- und netzwerkbasiertes Assessment muss für eine sehr flexible Gestaltung von Nutzer- und Administratorrechten offen sein, um derartigen unterschiedlichen Anwendungskontexten und hierarchischen organisatorischen Strukturen gerecht zu werden. Nicht zuletzt ist eine Offenheit des Systems wünschenswert, die eine automatische Anmeldung von Testteilnehmern auch über ein API (*Application Programming Interface*) erlaubt. Dadurch kann das System auch in Umgebungen eingesetzt werden, in denen die Identifikation von Testteilnehmern durch bereits bestehende Verfahren sichergestellt wird.

Datensicherheit

Für die Anwendung computerbasierten Assessments in Netzwerkumgebungen, insbesondere wenn Teile der Datenübertragung über das Internet erfolgen, ist Datensicherheit ein unverzichtbares Qualitätskriterium.

Datensicherheit bezieht sich insbesondere hinsichtlich *personenbezogener Daten* auf Sicherheit im Sinne des Schutzes vor unbefugtem Zugriff. Es muss möglich sein, durch Verschlüsselungstechniken und Passwortschutz die Sicherheit von durch das System gespeicherten Daten in einer Weise zu gewährleisten, die den aktuellen rechtlichen Anforderungen der Datenschutzgesetze des Bundes und der Länder gerecht wird. Wenn Testung und Datenspeicherung lokal auf einem einzelnen Computer erfolgen (vgl. Anwendungsszenarien in Kapitel 6), müssen die während der Testung gespeicherten Daten vor dem Zugriff von Dritten, die Zugriff auf die Festplatte dieses Computers haben, geschützt werden können. Um einen rechtsverbindlichen Schutz personenbezogener Daten sicherzustellen, darf das Kopieren und Einsehen dieser Daten für Unbefugte nur durch mutwillige Umgehung eines Kopier- und oder Passwortschutzes möglich sein.

Genauso wie der Schutz personenbezogener Daten sind auch Schutzmöglichkeiten für *Test- und Aufgabeninhalte* zu berücksichtigen. Hier spielt zum einen der Schutz des Urheberrechts eine Rolle – eine Vervielfältigung von Aufgabeninhalten sollte auch bei einer öffentlich zugänglichen Testdarbietung im Internet möglichst verhindert oder zumindest erschwert werden können. Bei der Verwendung von Tests im Kontext des System-Monitoring (vgl. Anwendungsszenarien in Kapitel 6) ist die Sicherheit von Aufgabeninhalten von besonderer zusätzlicher Bedeutung, weil öffentlich gemachte Aufgaben nicht mehr sinnvoll zu Evaluationszwecken eingesetzt werden können. Die Herausforderung bei der Entwicklung eines Basissystems für computer- und netzwerkbasiertes Assessment besteht darin, dass bei einer internetbasierten Testvorgabe auf den Com-

putern, an denen die Tests beantwortet werden das dauerhafte lokale Speichern von Test- und Aufgabeninhalten über die Testung hinaus weitgehend unmöglich sein sollte. Ein vollständiges Verhindern jeder Möglichkeit des Kopierens von Aufgabeninhalten (etwa durch Screenshots) ist allerdings genauso wenig zu gewährleisten wie ein hundertprozentiger Schutz vor dem unerlaubten Kopieren von Testinhalten bei einer Papier-Bleistift-Vorgabe.

Eine besondere Anforderung an die Sicherheit von Daten bei der Testdarbietung per Internet entsteht im Falle einer Testunterbrechung. Es ist wünschenswert, dass bereits erhobene Daten im Falle einer Trennung der Verbindung oder eines Computerabsturzes nicht verloren gehen. Es sollte möglich sein, eine Bearbeitung des Tests nach einem Abbruch an derselben Stelle wieder aufzunehmen. Hierzu ist eine lokale Datenspeicherung während der Testung notwendig, die eine Identifikation der getesteten Person bzw. des Computers gegenüber dem Server erlaubt. Ein Basissystem für computer- und netzwerkbasiertes Assessment sollte die technischen Voraussetzungen für eine derartige Zwischenspeicherung von Authentifizierungs- und/oder Testdaten bei gleichzeitiger Berücksichtigung der o.g. Überlegungen zum Datenschutz gewährleisten.

5.3.2 Spezifische diagnostische Anforderungen

Berücksichtigung von Aufgabenhierarchien

Verschiedene Aufgabenformate in pädagogisch-psychologischen Tests haben eine natürliche Hierarchie von Aufgabeninhalt und den innerhalb der Aufgabe zu beantwortenden Items. Beispiele sind einzelne Lücken innerhalb eines Lückentextes (z.B. C-Tests) oder verschiedene einzelne Fragen zu einem Text. Derartige strukturierte Aufgaben (im englischen auch *Testlets*) können nach einer empirischen Erprobung und Ermittlung der Itemeigenschaften nur als Ganzes vorgegeben werden, da sich bei einer Darbietung von Teilen die psychometrischen Eigenschaften der verbleibenden Items ändern können. Hierarchische Beziehungen von Aufgaben können für die Testvorgabe auch dann relevant sein, wenn mehrere Aufgaben nach demselben Prinzip konstruiert wurden und untereinander hierdurch zu ähnlich sind um derselben Person innerhalb einer Testung vorgelegt zu werden. Ein Basissystem für computer- und netzwerkbasiertes Assessment muss derartige Hierarchien bei der Verwaltung von Aufgaben, Items und Antworten sowie in Auswertungsalgorithmen berücksichtigen können.

Einsatz unterschiedlicher Auswertungsalgorithmen

In bestimmten Anwendungskontexten ist eine Ergebnisrückmeldung während oder unmittelbar nach der Testdurchführung wünschenswert; dies setzt eine Testauswertung durch das Testsystem voraus. Für die computerbasierte Auswertung eines Tests sollte ein Basissystem für computer- und netzwerkbasiertes Assessment die Implementierung verschiedener Auswertungsalgorithmen erlauben, auch Schätzverfahren für Personenparameter aus der Item-Response-

Theorie (IRT). Derartige Schätzungen setzen die Verwendung von Aufgabenparametern aus früheren Testerprobungen voraus, auf die das System dementsprechend während der Testung zugreifen muss. Eine IRT-basierte Schätzung von Personenparametern ist auch eine notwendige Voraussetzung für adaptives Testen (s.u.). Es sollte auch möglich sein, während einer Testung mehrere Testwerte parallel zu ermitteln und verwalten, wenn z.B. Aufgaben aus verschiedenen Bereichen gemischt dargeboten werden.

Zu Rückmeldezwecken sollte es auch möglich sein, bei der Testwertermittlung auf Personen- oder Gruppendaten zurückzugreifen. Dies ist z.B. notwendig, wenn sich Ergebnisse auf eine spezifische soziale Bezugsnorm beziehen sollen oder adjustierte Gruppenergebnisse zurückgemeldet werden sollen (z.B. Nachtigall, Kröhne, Enders & Steyer, im Druck).

Dynamische Kontrolle der Aufgabendarbietung

Das System sollte es ermöglichen, die Auswahl der Testaufgaben für jede Person unterschiedlich zu gestalten und diese Auswahl während der Testung nach bestimmten Kriterien dynamisch zu steuern. Das oben bereits erwähnte adaptive Testen ist eine mögliche Anwendung einer dynamischen Aufgabenauswahl, bei der auf die aktuelle Schätzung der Personenfähigkeit zurückgegriffen wird (z.B. Eggen, im Druck). Eine einfachere Variante des adaptiven Testens, die ebenfalls eine dynamische Aufgabenauswahl erfordert, ist verzweigtes Testen (*branched testing*), bei dem nach Bearbeitung und Auswertung eines ersten, für alle Personen gleichen Testteils aufgrund der gezeigten Leistung entschieden wird, welchen Teiltest eine einzelne Person als nächstes bearbeiten soll.

Insbesondere im Kontext von Large-Scale-Assessments gibt es jedoch auch andere Anwendungen für eine dynamische Aufgabenauswahl, z.B. Matrix-Designs, in denen jede Person nur eine bestimmte Teilmenge von Aufgaben bearbeitet. Während derartige Designs bei einer Papier-Bleistift-Vorgabe das Erstellen unterschiedlicher Testversionen erfordern und daher nur begrenzte Mengen von Versionen realisiert werden können, eröffnet computerbasiertes Assessment hier neue Möglichkeiten. Die Auswahl der Aufgaben kann auf Basis komplexer Versuchspläne erfolgen (z.B. Frey, Carstensen & Hartig, 2006) oder sogar von während der Testung gesammelten Informationen wie Itemexposition oder der Messgenauigkeit in Personengruppen Gebrauch machen (z.B. Wolf, 2006). Für solche Anwendungen sollte das Erstellen und Verwalten von *Testdesigns* mit spezifischen, ggf. dynamischen Vorgabekriterien innerhalb des Testsystems möglich sein. Dynamische Algorithmen zur Aufgabenauswahl müssen auch die oben bereits erwähnten möglichen hierarchischen Strukturen von Aufgaben berücksichtigen können, die es notwendig machen können, bestimmte Items immer oder nie gemeinsam darzubieten.

Technische Grundlagen für eine benutzerfreundliche Ergebnisrückmeldung

Wenn das hier skizzierte Basissystem auch für die Rückmeldung von Testergebnissen Verwendung finden soll, sollten diesbezügliche Anforderungen bei der Entwicklung bereits berücksichtigt werden. Spezifische Darbietungsstruktu-

ren mit Zugriff auf die entsprechenden Daten sollten sowohl für individuelle als auch gruppenbezogene Ergebnisse berücksichtigt werden, so dass es z.B. möglich ist, parallel zu einem Test eine angemessene Rückmeldung zu konstruieren. Auch das Erstellen allgemeiner Rückmeldevorlagen mit festen Elementen (z.B. Aufgabenbeispielen), die an spezifische Tests angepasst werden können, ist denkbar.

Einbezug menschlicher Beurteiler

Das System sollte die Möglichkeit berücksichtigen, dass eine unmittelbare computerbasierte Auswertung nicht für alle Aufgabenformate zweckmäßig ist. So können z.B. offene schriftliche oder gesprochene Antworten erfasst werden, die im Anschluss durch menschliche Beurteiler bewertet werden. Ein flexibles System für computer- und netzwerkbasiertes Assessment kann hier hilfreiche Strukturen zur Durchführung und Auswertung solcher Auswertungsprozesse zur Verfügung stellen. So könnten Beurteiler als eigene Nutzergruppe mit einer eigenen Benutzeroberfläche auf eine Datenbank mit offenen Antworten zugreifen und dort ihre Bewertungen abgeben. Über das System könnte hierbei eine optimale Verteilung der zu beurteilenden Aufgaben auf die Beurteiler sichergestellt werden.

Die so entstehenden Daten können unmittelbar genutzt werden, um die Übereinstimmungsgenauigkeit auf Beurteiler- und auf Aufgabenebene einzuschätzen und ggf. entsprechende Maßnahmen abzuleiten, z.B. das Nachschulen einzelner Beurteiler oder das Nachbessern der Kodieranweisungen für einzelne Aufgaben. Schließlich sollte es möglich sein, Auswertungsalgorithmen für die Beurteilungen zu implementieren, z.B. IRT-basierte Facettenmodelle mit Berücksichtigung von Beurteilereffekten. Die in einem derartigen Kontext gewonnenen Daten über die Beurteiler könnten ggf. für spätere Verwendungen weiter gespeichert und verwaltet werden (Beurteiler-Datenbanken).

Anpassung an Lehrkontexte

Ein Einsatz von computerbasierten Tests erfolgt häufig auch eingebunden in Lernprozesse, in denen Lernende z.B. ihren eigenen Kenntnisstand evaluieren können. Für derartige Anwendungen sollte es möglich sein, didaktisch geeignet aufbereitete Aufgabenlösungen zusammen mit Testaufgaben zu verwalten und die Testvorgabe so zu gestalten, dass getestete Personen während der Testung auf Lösungen zugreifen können.

Vom Nutzer anpassbare Darbietungs-Oberfläche

Je nach Anwendungskontext kann es wünschenswert sein, die Bearbeitung eines Tests auch für Personen mit körperlichen Einschränkungen (z.B. Sehbehinderungen) zu ermöglichen. Zumindest für einfache textbasierte Aufgabeninhalte sollte es möglich sein, eine Anpassung an derartige Bedürfnisse vorzusehen, wie sie unter dem Stichwort „barrierefreies Internet“ für Internet-Inhalte diskutiert werden. Grundsätzlich sollte bei der Entwicklung eines Basissystems

für computer- und netzwerkbasiertes Testen geprüft werden, ob Iteminhalte und technische Formen der Itemdarbietung getrennt verwaltet werden können.

Standards für Test- und Aufgabendokumentationen

Zuletzt soll hier noch einmal hervorgehoben werden, dass ein allgemeines System für computer- und netzwerkbasiertes Assessment, wenn es von einer breiten Anwendergemeinde verwendet wird, auch einen Beitrag zur wissenschaftsinternen Kommunikation leisten kann. Eine datenbankbasierte Verwaltung von Aufgaben- und Personendaten legt es nahe, Standards für die Struktur dieser Daten zu entwickeln. So könnten bestimmte Informationen über Tests und Aufgaben, wie z.B. der Status der empirischen Erprobung, zu Pflichtinformationen erklärt werden. Metadaten für Items sind nicht nur zu Dokumentationszwecken notwendig, sondern auch für die Zusammenstellung von Tests unentbehrlich. Eine Herausforderung bei der Dokumentation von Items ergibt sich daraus, dass auch Itemparameter (basierend auf klassischer Testtheorie oder IRT) gespeichert werden müssen, die sich jedoch immer auf den Einsatz eines Items im Kontext eines spezifischen Tests und auf eine spezifische Stichprobe beziehen. Die Frage nach der Struktur und Organisation der Itemdokumentation ist jedoch nicht nur eine technische Anforderung an ein mögliches Testsystem, sondern auch an die Organisation der Nutzergemeinde.

5.4 Anforderungen computer- und netzwerkbasierten Assessments an Bildungsinstitutionen

Um computerbasiertes Assessment beispielsweise in Schulen durchführen zu können, müssen auch dort zahlreiche Voraussetzungen erfüllt sein. Neben dem zusätzlichen Personalaufwand, der zumindest Anfangs in einem Mehraufwand für das Lehrpersonal besteht, welches die Vorgabe zentraler computerbasierter Leistungstests durchführt und überwacht, sollten außerdem für die Schulen zuständige IT-Fachkräfte verfügbar sein, die bei möglichen Problemen, Ausfällen, etc. sofort Abhilfe schaffen können. Dies ist vor allem im Falle von so genanntem *High Stakes Testing* unumgänglich, bei dem die Testergebnisse unmittelbare Konsequenzen wie zum Beispiel Notenvergabe oder Selektionsentscheidungen haben. Hier dürfen keinesfalls technische Probleme zu Nachteilen für betroffene Schülern führen. Ferner müsste es in jeder teilnehmenden Schule eine Grundausstattung im IT-Bereich geben, die genügend Computer mit einer ausreichenden technischen Ausstattung, sowie einen den jeweiligen Zwecken genügenden, stabilen Internetzugang beinhaltet.

Aus dem Bericht des Bundesministeriums für Bildung und Forschung (BMBF) zur IT-Ausstattung der allgemein bildenden und berufsbildenden Schulen in Deutschland im Jahre 2004 (Quelle) gehen einige Hinweise bezüglich der Erfüllung der Voraussetzungen seitens der Bildungsinstitutionen für die Anwendung von computerbasiertem Assessment hervor.

So ist das Verhältnis Schüler-Computer in den allgemein bildenden Schulen in der Sekundarstufe I und II 13:1 und bei den berufsbildenden Schulen 15:1.

Auch in Grundschulen kommt auf 15 Schüler im Schnitt ein Computer. Insgesamt sind 98% der Schulen mit stationären oder mobilen, für den Einsatz im Unterricht verwendbaren Rechnern ausgestattet. Insgesamt hat sich die Anzahl der in Schulen vorhandenen Computer zwischen 2001 und 2004 von 450.000 auf 950.000 mehr als verdoppelt.

Die Anzahl von Computern mit Internetverbindung verfünffachte sich vom Jahre 2001 bis 2004 auf 550.000, was 68% der in der Schule verwendeten Computer entspricht. 78% der Rechner sind außerdem innerhalb der Schule an ein lokales Netzwerk angeschlossen.

In Grundschulen sind ferner 43% der Computer multimediatauglich, in der Sekundarstufe I und II sogar 69%.

Diese Zahlen entsprechen zwar den im Aktionsplan „eLearning" der Europäischen Kommission ausgesprochenen Empfehlungen. Allerdings dürfte dies kaum genügen, um beispielsweise ein flächendeckendes Testen im Klassenverband durchzuführen, zumal es sich bei den Zahlen um Durchschnittswerte handelt und es daher möglich ist, dass die Anzahl der Computer pro Schüler in einigen Schulen deutlich niedriger ausfällt.

Da dieser Bericht bereits 2004 verfasst wurde ist zu hoffen, dass sich die Situation bezüglich der IT-Ausstattung deutscher Schulen inzwischen noch weiter verbessert hat, so dass dort möglicherweise computerbasiertes Assessment mittlerweile auch im Klassenverband durchführbar wäre.

Im Bereich der Erwachsenenbildung sind vermutlich andere Voraussetzungen zu beachten: Dort findet institutionalisierte Bildung in einem sehr viel geringeren Maße statt. Gerade aus diesen Gründen ist es hier umso wichtiger, dass geprüft wird, ob Zugangsmöglichkeiten für alle Interessenten zu entsprechenden Tests, beispielsweise im Bereich der zertifizierten Weiterbildung, vorhanden sind. Eine Möglichkeit um dies zu Gewährleisten wäre die Einrichtung von Testzentren, etwa in Volkshochschulen oder Gemeindezentren, die in den meisten Städten vorhanden und gut erreichbar sind. Dies würde allerdings erhebliche finanzielle Belastungen bedeuten, und die Frage der Finanzierung und der Verantwortlichkeit bliebe zu klären. Ein weiterer und vermutlich auf Dauer besser gehbarer Weg ist das Testen via Internet. Dafür wiederum müssten jedoch gleichfalls entsprechende Voraussetzungen erfüllt sein was bedeutet, dass möglichst alle oder zumindest ein Großteil der in Deutschland lebenden Bevölkerung einen Haushalt mit Internet- Zugang besitzen. Die ARD/ZDF Online-Studie 2006 (Quelle) berichtet, dass 38,6 Millionen Erwachsene, d.h. ab 14 Jahren, regelmäßig das Internet nutzen. Dies entspricht 59,5% der Bevölkerung Deutschlands.

Dies zeigt, dass obwohl der Anteil der Personen mit Internetzugang noch immer moderat anwächst, hier noch kein Zugang für alle gegeben ist. Daher müssen auch für das Testen via Internet Testzentren für die Personen ohne eigenen Internetzugang vorhanden und erreichbar sein.

6 Anwendungsszenarien computer- und netzwerkbasierten Assessments

Astrid Jurecka & Johannes Hartig

Im Folgenden sollen einige mögliche Anwendungsszenarien eines computer- und netzwerkbasierten Assessment-Systems, wie es im vorangegangenen Kapitel skizziert wurde, ausgeführt werden. Diese Szenarien stellen eine Auswahl unterschiedlicher Anwendungskontexte dar, die der Illustration der mit einem solchen System denkbaren vielfältigen Möglichkeiten dient. Es wird hierbei weder der Anspruch auf Vollständigkeit der Einsatzmöglichkeiten erhoben, noch darauf, dass ein Einsatz computer- und netzwerkbasierten Assessments in den hier beschriebenen Kontexten nur in der hier skizzierten Weise möglich wäre. Zu jedem der hier aufgeführten Anwendunskontexte sind verschiedene alternative Möglichkeiten einer Implementierung computer- und netzwerkbasierten Assessments denkbar, die sich von den hier skizzierten in Einzelheiten oder grundlegend unterscheiden können. Es ist für die folgende Darstellung der Anwendungsszenarien auch sekundär, wie Inhalt und Format der Aufgaben definiert werden. Vorausgesetzt wird aber, dass die Aufgabenerstellung und -bearbeitung innerhalb des Systems vorgenommen werden kann.

Für jedes Anwendungsszenario werden in Kürze die notwendige *Server- Clientstruktur*, die benötigten *Nutzerprofile* sowie die *Testsituation* beschrieben. Hinsichtlich der Nutzerprofile werden hier lediglich die von Seiten des Systems notwendigen Rechte aufgeführt, über die unterschiedliche Personen im jeweiligen Anwendungskontext verfügen müssen. Diese Personen, die in den hier skizzierten Szenarien teilweise sehr unterschiedliche Rollen und Funktionen haben, benötigen bei einer Realisierung entsprechende, auf diese Rollen zugeschnittene, Benutzeroberflächen. Auf die Inhalte und Gestaltung dieser Oberflächen wird hier nicht eingegangen. Es ist jedoch festzuhalten, dass neben den im vorangegangenen Kapiteln genannten Anforderungen auch das flexible Zusammenstellen von funktionsspezifischen, anwenderfreundlichem Oberflächen für Nutzer und Administratoren mit unterschiedlichen Rechten eine wichtige Anforderung an ein Basissystem für computer- und netzwerkbasiertes Assessment darstellt.

6.1 Anwendung in Bildungsevaluation und Lehre

6.1.1 Aufgabenentwicklung

Ein netzwerkbasiertes System zur Aufgabenverwaltung, das einen Zugriff auf Datenbanken auch per Internet erlaubt, kann nicht nur zur Testvorgabe verwendet werden. Ein Einsatz eines derartigen Systems bietet bereits bei der Aufgabenentwicklung interessante Möglichkeiten, insbesondere wenn diese in einem interdisziplinären, auf verschiedene räumliche Standorte verteilten Entwicklerteam erfolgt. In einem derartigen Szenario können Aufgabeninhalte,

Antwortmöglichkeiten und Bewertungskriterien in einer Datenbank abgelegt werden und mit der Möglichkeit versehen werden, Kommentare und alternative Varianten abzulegen. Durch ein derartiges System kann die Bearbeitung und Revision durch verschiedene Personen zeitlich asynchron erfolgen, die zu einer Aufgabe abgelegten Informationen können von allen Entwicklern jederzeit eingesehen werden. Eine zentrale Instanz kann die Supervision des Entwicklungsprozesses übernehmen und entscheiden, wann die Arbeit an einer Aufgabe als abgeschlossen betrachtet werden kann und welche der vorgeschlagenen Alternativen dann die endgültige Version darstellt. Die so entwickelten Aufgaben können dann aus dem System heraus direkt für eine empirische Erprobung weiterverwendet werden.

Server-Clientstruktur

Ein derartiges Szenario kann mit einem zentralen Server realisiert werden, auf dem die Aufgabeninhalte einschließlich alternativer Versionen für einzelne Bestandteile sowie Kommentare der Entwickler gespeichert sind. Die Entwickler greifen per Internet- oder LAN-Verbindung auf diesen Server zu, um neue Aufgaben einzustellen oder bereits vorhandene zu bearbeiten oder zu kommentieren. Der Zugriff durch die supervidierende Person kann ebenfalls über eine Netzwerkanbindung von einem beliebigen Arbeitsplatz erfolgen.

Nutzerprofile

- **Aufgabenentwickler:** Die Entwickler haben in diesem Szenario Rechte, die in anderen Szenarien administrativen Rechten entsprechen. Sie können Aufgaben erstellen und bestehende Aufgaben durch das Einfügen alternativer Inhalte bearbeiten. Ein Entwickler kann alle Informationen zu einer Aufgabe einsehen, jedoch keine Änderungen an von anderen Entwicklern eingestellten Informationen vornehmen.
- **Systemadministrator (Supervisor):** Bei einer zentralen Institution, die für die Aufgabenentwicklung verantwortlich ist (z.B. Landesqualitätsagentur) liegen die vollen Zugriffsrechte auf das System, einschließlich der Bearbeitung und Löschung von durch die Entwickler eingestellten Aufgabeninhalten und Kommentaren. Die zentrale Administration hat auch die Möglichkeit, die Bearbeitung einer Aufgabe zu beenden und diese für die Bearbeitung zu sperren.

6.1.2 System-Monitoring im Bildungswesen

Wenn allgemeingültige Curricula und Bildungsstandards in verschiedenen Schulfächern regelmäßig auf ihre Erfüllung hin überprüft werden sollen, könnte eine teilweise oder vollständige Durchführung der notwendigen Tests mit Hilfe eines netzwerkbasierten Testsystems erfolgen. Tests könnten hierbei regelmäßig zu festen Zeitpunkten (z.B. am Ende des Schuljahres) oder an verschiedenen Schulen zu zufälligen Zeitpunkten stattfinden. *Testinhalte* würden jedenfalls zentral entwickelt und zur Verfügung gestellt, z.B. basierend auf den jeweiligen

Curricula eines Bundeslandes oder auf nationalen Bildungsstandards. Gleichzeitig mit den Leistungstests wäre auch die Vorgabe von Schüler- und Lehrerfragebögen vorstellbar. Im Rahmen eines System-Monitorings ist es auch möglich, Ergebnisse auf Schulebene an die Schulen zurückzumelden.

Server-Clientstruktur

Unter der Annahme, dass eine Breitband-Internetanbindung aller Schulen (noch) nicht vorausgesetzt werden kann, ist die Realisierung einer groß angelegten, regelmäßigen Durchführung von Leistungstests an Schulen mit Hilfe einer hierarchischen Serverstruktur denkbar. Die Gesamtheit der Aufgaben kann in diesem Szenario auf einem zentralen, für authentifizierte Nutzer per Internet zugänglichen Server gespeichert und verwaltet werden. An jeder Schule wird ein Server innerhalb eines lokalen Netzwerks eingerichtet, der die notwendigen Aufgaben und Testdurchführungsanweisungen vom zentralen Server bezieht und auf dem die Aufgaben lokal gespeichert werden. Die Definition und Einrichtung der teilnehmenden Schülerstichprobe wird auf dem lokalen Server nach zentral vorgegebenen Kriterien (z.B. alle siebten Klassen) vorgenommen. Die Testvorgabe erfolgt auf Einzelarbeitsplätzen, z.B. in einem Computerpool, die auf den lokalen Server zugreifen, wobei die Aufgabenauswahl (z.B. für adaptive Tests oder im Rahmen eines Matrix-Designs) durch den lokalen Server gesteuert wird. Die Speicherung der erfassten Antworten erfolgt auf dem lokalen Server. Die Testdurchführung kann innerhalb jeder Schule wenn nötig zeitlich verteilt organisiert werden, so dass z.B. ein Computerpool zur sukzessiven Testung mehrerer Klassen verwendet werden kann. Nach dem Abschluss der Testdurchführung können die lokalen Server jeder Schule die erfassten Daten an den zentralen Server übermitteln, auf dem die Ergebnisse aller Schulen zusammengeführt und ausgewertet werden.

Nutzerprofile

- **Testteilnehmer (Schüler):** Für die getesteten Schüler werden Nutzerkonten eingerichtet, mit denen lediglich eine Anmeldung am System und die Bearbeitung der vorgegebenen Aufgaben möglich ist. Soweit Datenschutzrechtlich möglich, kann die Identifikation der Schüler auch dazu dienen, die Testdaten mit anderen personenbezogenen Daten (z.B. Alter, Geschlecht) in Beziehung zu setzen und diese bei der Auswertung zu nutzen. Es ist nicht notwendig, dass die Schüler ihren Nutzernamen und ihr Passwort kennen, die Anmeldung kann durch den Testleiter erfolgen (s.u.).
- **Testleiter (z.B. Klassenlehrer):** Die für die Durchführung der Testung in einer Klasse verantwortliche Person muss sicherstellen, dass jeder Schüler mit dem richtigen Nutzernamen angemeldet wird. Hierzu sollte sie über Rechte verfügen, die Nutzerdaten der Schüler innerhalb einer Klasse zu verwalten, auch um ggf. Schüler hinzufügen. Falls rechtlich möglich, könnte ein Klassenlehrer auch zusätzliche Informationen zu den Schülern in das System eingeben, z.B. Alter, Geschlecht oder Schulnoten. Im hier skizzierten Szena-

rio haben Testleiter keinen Zugriff auf Aufgabeninhalte und keinen Einfluss auf die Aufgabenauswahl.

- **Lokale Administratoren:** In jeder Schule sollte ein qualifizierter Verantwortlicher für die gesamte Testdurchführung verfügbar sein. Diese Person hat administrativen Zugriff auf die Nutzerdaten der Schüler und der Testleiter sowie die nötigen Rechte, um die Datenübertragung zwischen dem lokalen Server und dem zentralen Server zu steuern (d.h. das Beziehen der Aufgabeninhalte und die Übermittlung der erfassten Daten). Für die lokalen Administratoren ist ein Zugriff auf die Aufgabeninhalte und die Aufgabenauswahl nicht notwendig, könnte aber aus Akzeptanzgründen in Betracht gezogen werden. Die Rolle lokaler Administratoren könnte, soweit kein entsprechend qualifiziertes Personal an Schulen vorhanden ist, auch von externen Arbeitskräften (z.B. von Landesqualitätsagenturen) übernommen werden. Falls eine Schulevaluation gewünscht wird, hat der lokale Administrator ferner Zugriff auf Untersuchungsergebnisse auf Schulebene.
- **Systemadministratoren:** Bei der durchführenden zentralen Institution (z.B. Landesqualitätsagentur) liegen die vollen Zugriffsrechte auf das System, einschließlich der Verwaltung und Bearbeitung von Aufgabeninhalten und Aufgabenauswahl-Algorithmen und dem Zugriff auf die Gesamtheit der erfassten Daten zum Zwecke der Auswertung.

Testsituation

Die Testung findet in diesem Szenario in Klassen bzw. Gruppen statt. Testort kann beispielsweise ein Computerraum sein. Während der Bearbeitung der Aufgaben muss ein Testleiter anwesend sein, der zu Beginn sicherstellt, dass jeder Testteilnehmer mit den richtigen Nutzerdaten angemeldet ist. Der Testleiter ist verantwortlich für die instruktionsgemäße Bearbeitung der Tests und soll eine lokale Speicherung oder Kopie von Aufgabeninhalten verhindern. Für einfache technische Fragen steht der Testleiter direkt zur Verfügung, bei größeren technischen Schwierigkeiten sollte der lokale Administrator der Schule erreichbar sein.

6.1.3 Unterrichtsevaluation und -entwicklung

Für eine Überprüfung der Effektivität des eigenen Unterrichts durch die Lehrperson selbst sowie für die Unterrichtsentwicklung kann ein netzwerkbasiertes Testsystem ebenfalls eine hilfreiche Rolle spielen.

In einem Szenario zum Einsatz eines computer- und netzwerkbasierten Assessmentsystems in der Unterrichtsevaluation und -entwicklung können Aufgaben von zentraler Stelle zur Verfügung gestellt werden, die auf den jeweils gültigen Curricula basieren. Lehrer haben dann über das Internet Zugriff auf einen Aufgabenpool und können sich die zum jeweiligen Unterrichtsthema passenden Aufgaben heraussuchen. Diese können dann den Schülern einer Klasse oder eines Kurses im Rahmen einer Testung oder auch als Teil einer Klassenarbeit vorgegeben werden, wenn diese z.B. in einem Computerpool durchgeführt werden. Die Auswertung der Aufgaben erfolgt automatisch über das System. Lehrer

hätten somit die Möglichkeit, unmittelbar einen Überblick über den momentanen Leistungsstand der Klasse hinsichtlich bestimmter Themen zu erhalten. Ferner wäre durch eine Verwendung dieser Aufgaben ein Abgleich des Leistungsstands der Klasse mit dem in den Curricula erwünschten Output oder ggf. mit sozialen Bezugsnormen möglich. Sollte die tatsächliche Leistung nicht der erwünschten entsprechen, haben Lehrer die Möglichkeit, darauf mit verstärkten Übungsmaßnahmen oder einer Umstellung des Unterrichtsplans relativ schnell und flexibel zu reagieren. Bei dieser Form der Unterrichtsevaluation durch die Lehrperson selbst handelt es sich um eine freiwillige Maßnahme.

Die Mittel und Materialien für eine solche Form der Evaluation zur Verfügung zu stellen, könnte einen wichtigen Beitrag zur Umsetzung von Bildungsstandards und Curricula bringen. Ferner kann diese Art der Unterrichtsevaluation Lehrern und Schülern größere Sicherheit bezüglich ihres aktuellen Leistungsstandes geben, vor allem auch in Bezug auf die in den nächsten Jahren vermehrt anstehenden, zentralen und auf Curricula und Bildungsstandards basierenden Leistungsüberprüfungen.

Server-Clientstruktur

Eine Realisierung dieses Szenarios ist durch die Einrichtung eines zentralen Servers (z.B. an Landesqualitätsagenturen) denkbar, auf dem Aufgaben bereitgestellt werden. Per Internet erfolgt in regelmäßigen Abständen eine Synchronisierung der zentral gespeicherten Aufgaben mit einem lokalen Server eines schulinternen Netzwerks. Von diesem Server können seitens der dem Netzwerk angeschlossenen Einzelarbeitsplätze die für die Testvorgabe notwendigen Informationen und Aufgaben bezogen werden, Lehrer greifen zum Zweck der Aufgabenauswahl und -zusammenstellung auf den lokalen Server zu. Nach Bearbeitung der Aufgaben werden die Ergebnisse der Testung von den Einzelarbeitsplätzen an den lokalen Server weitergegeben und dort auf Individual- und Klassenebene gespeichert und ausgewertet.

Nutzerprofile

- **Schüler**: Auch hier werden für die Schüler Nutzerkonten eingerichtet, die ihnen die Bearbeitung der Aufgaben ermöglichen. Die Schüler melden sich entweder selbst an, oder werden vom Testleiter mit einem Passwort im System angemeldet. Dadurch kann der einzelne Schüler später identifiziert und die Testergebnisse können einzeln ausgewertet und zurückgemeldet werden.
- **Lehrer:** Um die Aufgaben an erster Stelle erhalten und auswählen zu können, muss dem Lehrer das Recht erteilt werden, auf den zentralen Server Zugriff zu haben, auf dem der Aufgabenpool gespeichert ist. Der Lehrer meldet sich mit einem Passwort an und erhält dann Zugriff auf die Aufgaben, die er via Internet herunterladen kann, um sie dann auf dem Server des lokalen Netzwerks zu speichern. Der Lehrer hat innerhalb des lokalen Netzwerkes das Recht, Aufgaben in das Netzwerk zu stellen und diese auszuwählen sowie auf die Ergebnisse zuzugreifen und diese auszuwerten. In diesem Fall ist der Lehrer übrigens die einzige Person, die Zugriff auf die

Ergebnisse der Testung hat. Ferner hat er das Recht, die Nutzerdaten innerhalb seiner Klasse oder seines Kurses zu verwalten.

- **Lokale Administratoren:** In jeder Schule ist ein qualifizierter Administrator für die Synchronisierung des lokalen Servers mit dem zentralen Server verantwortlich. Neben den zur Steuerung dieser Datenübertragung notwendigen Rechten hat diese Person zu lokalen Administrationszwecken Zugriff auf die Nutzerdaten der Lehrer, nicht jedoch auf die individuellen oder klassenspezifischen Testergebnisse.
- **Systemadministratoren:** In diesem Szenario beschränkt sich das Zugriffsrecht der zentralen Institution (z.B. Landesqualitätsagentur) auf das Verwalten und Bearbeiten von Aufgaben und Aufgabeninhalten. Sie hat kein Recht, auf die Testergebnisse zuzugreifen.

Testsituation

Auch hier wäre eine Bearbeitung in einem Computerpool einer Schule denkbar. Der Klassen- oder Kurslehrer ist als Aufsichtsperson und Testleiter anwesend. Er stellt sicher, dass alle Schüler korrekt im System angemeldet sind, und steht für mögliche Fragen und zur Lösung kleinerer technischer Probleme zur Verfügung.

6.1.4 Hochschullehre

Die Verwendung eines Basissystems für computerbasiertes Assessment ist auch im Bereich der Hochschullehre gut vorstellbar. Die Verwendung wäre sowohl bei gängigen Formen von Präsenzveranstaltungen als auch bei Fernuniversitäten sinnvoll.

Denkbar wäre hier beispielsweise eine Entwicklung von Aufgaben zu bestimmten Themengebieten. Diese könnten den Studenten zu Übungszwecken oder zur Prüfungsvorbereitung durch den Lehrenden zugänglich gemacht werden, indem die Aufgaben auf einen zentralen, mit dem Internet verbundenen Server gestellt werden. Die Studierenden können nun von zu Hause via Internet die Aufgaben bearbeiten und erhalten eine sofortige Ergebnisrückmeldung vom System, welches die Aufgaben automatisch auswertet. Somit könnten die Aufgaben zur Überprüfung und Verbesserung der eigenen Leistung dienen.

Sinnvoll könnte zu diesem Zweck auch beispielsweise die Entwicklung von Itemtypen sein, die Teillösungen und die komplette Lösung der Aufgabe bereits beinhalten. Vorstellbar wäre das zum Beispiel bei Mathematikaufgaben, in denen der Student zwischen verschiedenen Bearbeitungsmodi wählen kann, nämlich entweder die Aufgabe komplett alleine zu bearbeiten, oder sich zwischendurch Teillösungen und Rechenschritte anzeigen zu lassen, um somit den korrekten Lösungsweg mitverfolgen und sich aneignen zu können.

Die Aufgaben könnten entweder vom Seminarleiter selbst, oder von einem speziellen Fachgremium zur universitätsübergreifenden Verwendung entwickelt werden.

Nutzerprofile

- **Studenten:** Für jeden Studenten eines Seminars wird ein Nutzerkonto eingerichtet. Damit ist es ihnen möglich, sich mit einem Passwort im System anzumelden, die Aufgaben zu bearbeiten und eine Ergebnisrückmeldung zu erhalten.
- **Lehrende:** Diese haben das Recht, Aufgaben auszuwählen (und ggf. selbst zu entwickeln), zu verwalten und Nutzerkonten einzurichten. Sollten die Aufgaben an zentraler Stelle entwickelt worden sein, wird ihnen ferner das Recht erteilt, Zugriff auf den dortigen Server zu erhalten um sich die passenden Aufgaben auszuwählen und herunterzuladen. Da die Aufgaben zu Übungszwecken gedacht sind, hat der Lehrende keinen Zugriff auf die Ergebnisse der Aufgaben.
- **Systemadministrator:** Im Falle von zentral entwickelten Aufgaben hat hier der Systemadministrator das Recht, Aufgaben zu verwalten und zur Verfügung zu stellen.

Testsituation

Da es sich hier um Übungsaufgaben und somit eher um eine „Low stakes testing"-Situation handelt, hat primär der Student selbst ein Interesse daran, die Aufgaben zu bearbeiten. Daher dürfte die Motivation, bei der Beantwortung der Aufgaben zu betrügen, vermutlich ziemlich gering sein. Auch ist hier eine Geheimhaltung der Aufgaben an sich nicht zwingend notwendig. Bei der Aufgabenbearbeitung ist daher keine Aufsichtsperson anwesend.

6.2 Anwendung in der Forschung

6.2.1 Wissenschaftliche Large-Scale-Assessments in der schulischen Bildung

Computerbasiertes Assessment könnte auch eine Rolle bei Large-Scale-Studien im Bereich der schulischen Bildung spielen. Aufgaben in großen nationalen sowie internationalen Vergleichsstudien wie beispielsweise PISA oder TIMSS werden in diesem Szenario den teilnehmenden Schülern in computerisierter Form vorgegeben. Die in diesen Studien verwendeten Aufgaben werden vorher an zentraler Stelle nach den jeweiligen inhaltlichen Gesichtspunkten entwickelt. Die Aufgaben werden typischerweise in einem Matrix-Design vorgegeben, bei dem jeder Schüler nur einen Teil der Aufgaben bearbeitet; bei einer computerbasierten Testung ist auch adaptives Testen denkbar. Die Auswertung der Aufgaben erfolgt, soweit möglich, automatisch. Die Ergebnisse der Aufgabenformate, die einer menschlichen Beurteilung bedürfen, können später von den Auswertern in das System eingegeben werden.

Server-Clientstruktur

Auch in diesem Szenario ist die Realisierung einer solchen Studie mit Hilfe einer hierarchischen Serverstruktur vorstellbar. Um die Verfügbarkeit und Ver-

gleichbarkeit der nötigen Hardware sicherzustellen, erfolgt die Testung an im Rahmen der Studie bereitgestellten Notebooks, die in einem lokalen Netzwerk verbunden sind. Der Server des lokalen Netzwerks wird ebenfalls im Kontext der Studie gestellt und ist in einem lokalen Funknetz (*Wireless Local Area Network; WLAN*) mit den Computern, auf denen die Testbeantwortung erfolgt, verbunden. Der lokale Server und das WLAN werden jeweils ad hoc nur für die Testdurchführung eingerichtet. Die Ergebnisse werden auf dem lokalen Server gespeichert. Von dort wiederum werden die Daten per Internet an einen zentralen Server übermittelt, welcher durch die durchführende Institution eingerichtet wurde, und auf dem die Daten aller teilnehmenden Schulen und Schüler eines Landes gesammelt werden. Dort erfolgt im Falle einer internationalen Studie eine erste Auswertung der Ergebnisse, im Falle einer nationalen Studie die endgültige Auswertung.

Die Daten einer internationalen Studie werden an einen weiteren zentralen Server weitergegeben, auf dem die Daten der Schüler aller teilnehmenden Länder gesammelt werden. Dort erfolgt dann die endgültige Auswertung der Studie.

Die hier im Kontext eines wissenschaftlichen Large-Scale-Assessments skizzierte Systemkonfiguration wäre selbstverständlich auch für einen Einsatz im bundes- oder landesinternen System-Monitoring denkbar. Umgekehrt kann auch ein Large-Scale-Assessment auf vorhandene technische Infrastruktur an den Schulen zurückgreifen, wenn diese in hinreichendem Umfang vorhanden ist.

Nutzerprofile

- **Schüler:** Auch hier werden für die einzelnen Schüler Nutzerkonten eingerichtet. Nach erfolgter Anmeldung im System sind dann die Aufgaben zur Bearbeitung freigegeben.
- **Testleiter und lokale Administratoren:** Bei großen internationalen Studien unterstehen die verwendeten Aufgaben häufig hoher Geheimhaltung, da ein bekannt werden der Aufgaben vor Beendigung der Studie zu stark verfälschten Ergebnissen führen könnte. Aus diesem Grund wäre es in einem solchen Szenario denkbar, dass die Aufgaben den Schülern auf vom Testleiter mitgebrachten Notebooks vorgegeben würden. In diesem Fall kann es sich bei Testleiter und lokalem Administrator um ein und dieselbe Person handeln. Der lokale Administrator hat das Recht, Aufgaben auf dem Server des lokalen Netzwerks zu installieren und Benutzerkonten einzurichten.
- **Systemadministratoren national:** Bei der die Studie durchführenden zentralen Institution liegen die vollen Zugriffsrechte auf das nationale System und auf die Gesamtheit der national erfassten Daten zum Zwecke der Auswertung. Im Falle nationaler Vergleichsstudien liegt hier, je nach Auftraggeber und Forschungsziel der Studie, auch das Recht, Aufgabeninhalte zu bearbeiten und Aufgaben auszuwählen. Dies ist in internationalen Vergleichsstudien keineswegs der Fall.
- **Systemadministratoren international:** Bei dem zentralen Institut einer internationalen Studie liegen die Zugriffsrechte auf das komplette System.

Dies schließt hier auch die Verwaltung und Bearbeitung von Aufgabeninhalten und Aufgabenauswahl-Algorithmen und den Zugriff auf die Gesamtheit der international erfassten Daten zum Zwecke der Auswertung ein.

Testsituation

Während der Testung ist ein Testleiter anwesend, der einen korrekten Ablauf der Studie sicherstellt, Instruktionen gibt und bei Fragen sowie zur Lösung technischer Probleme zur Verfügung steht. Die Testung erfolgt auf für die Studie zur Verfügung gestellten Notebooks.

6.2.2 Laboruntersuchungen

Bei Laboruntersuchungen handelt es sich meist um Forschungsstudien zur Grundlagenforschung, die unter bestimmten experimentellen und kontrollierten Bedingungen stattfinden. Laboruntersuchungen können Inhalte verschiedenster Natur beinhalten. Denkbar ist hier beinahe jeder Inhaltsbereich der Pädagogik oder Psychologie. Auch kann es sich hierbei sowohl um die Untersuchung einer einzelnen Testperson handeln, als auch um eine kleinere Gruppe von Personen, die gleichzeitig an dem gleichen Experiment oder auch an einer gemeinsamen Gruppenaufgabe arbeiten.

Server-Clientstruktur

Im Falle der Untersuchungsdurchführung in Einzelsitzungen an einem einzigen Computer kann auf diesem ein virtueller Server installiert werden, der die Aufgabenauswahl und Datenspeicherung steuert. In diesem Fall wären Server und Client identisch, da die Aufgabendarbietung und -beantwortung auf demselben Computer erfolgt. Im Falle eines kleinen lokalen Netzwerkes, wie es häufig in Computerlaboren der Fall ist, würde einer der Computer als Server, und die restlichen als damit verbundene Einzelarbeitsplätze eingerichtet. Die Aufgaben werden vom Server zur Verfügung gestellt, und dort auch nach Bearbeitung der Aufgaben durch die Probanden die Ergebnisse gesammelt und ausgewertet.

Nutzerprofile

- **Probanden:** Für die an der Studie teilnehmenden Probanden werden ein Nutzerkonto und ein Passwort eingerichtet. Die Probanden können vom Versuchsleiter im System angemeldet werden, und dann die vorgegebenen Aufgaben bearbeiten.
- **Lokaler Administrator (Versuchsleiter):** Der lokale Administrator hat Zugriff auf das System, soweit dies für die Durchführung notwendig ist (z.B. das Auswählen von experimentellen Untersuchungsbedingungen) und kann Nutzerkonten einrichten und bearbeiten.
- **Systemadministrator (Untersuchungsleiter):** Die mit der Forschungsleitung betraute Person hat die Rechte, Aufgaben zu erstellen und zu bearbeiten, die Vorgabekriterien zu bearbeiten sowie vollständig auf die Ergebnisse zuzugreifen.

Testsituation

Üblicherweise ist bei Laboruntersuchungen ein Versuchsleiter anwesend, der die Instruktionen an die Probanden gibt, die Untersuchung startet und überwacht sowie bei eventuellen Fragen seitens der Probanden zur Verfügung steht.

6.2.3 Internetbasierte Grundlagenforschung

Das Internet wird in der psychologischen Forschung schon lange als ökonomisches Medium der Datenerhebung genutzt, über das sich einfach und schnell große und relativ heterogene Stichproben rekrutieren lassen. Die meisten internetbasierten psychologischen Untersuchungen sind Fragebogenuntersuchungen, wie sie vor allem für die Persönlichkeitsforschung typisch sind (z.B. Buchanan & Smith, 1999; Jude, Hartig & Rauch, 2005), es werden jedoch auch experimentelle Untersuchungen per Internet durchgeführt (z.B. Birnbaum, 2000; Hölzel, Hartig, Rabl & Moosbrugger, 2004; Reips, 2002).

Internetbasierte Untersuchungen sind in der psychologischen oder pädagogischen Forschung dann eine interessante Möglichkeit, wenn die Zusammensetzung der Stichprobe keine große Bedeutung hat und vor allem wenn es in Kauf genommen werden kann, dass die Untersuchungssituation nicht kontrolliert werden kann. Bei Fragebogenbeantwortung wird dies in der Regel auch bei Papier-Bleistift-Vorgabe als akzeptabel erachtet, bei Leistungstests kann dies unter Umständen als vertretbar erscheinen, wenn für die gestestete Person aus dem Testergebnis keine Konsequenzen resultieren und wenn die Untersuchungsteilnehmer an einer Rückmeldung ihrer Ergebnisse interessiert sind. Eine individuelle Ergebnisrückmeldung ist als Anreiz für die Teilnahme an internetbasierten psychologischen Unersuchungen von großer Bedeutung und reicht häufig schon allein aus, um recht große Teilnehmerzahlen zu rekrutieren – so nahmen an den von Hölzel et al. (2004) berichteten Studien in jeweils drei bis vier Wochen weit mehr als 1000 Personen teil, obwohl es außer einer Ergebnisrückmeldung keinerlei Anreiz zur Teilnahme gab.

Internetbasierte psychologische Untersuchungen – sowohl fragebogenbasiert, mit Leistungstests oder experimentellen Untersuchungsdesigns – sind auch mit einem Basissystem für computer- und netzwerkbasiertes Assessment durchführbar. Ein solches System würde für den Forschenden eine flexible Alternative zu von kommerziellen Anbietern verfügbaren Befragungsplattformen oder dem vollständig eigenen Durchführen der nötigen serverseitigen Programmierung darstellen. *Testinhalte* – im Folgenden wie in den anderen Szenarien auch als Aufgaben bezeichnet – müssen sich nicht auf Leistungsaufgaben im engeren Sinn beschränken. Es können auch computerbasierte Fragebögen oder beliebige experimentelle Stimuli mit geeigneten Reaktionsmöglichkeiten dargeboten werden.

Server-Clientstruktur

Im Szenario internetbasierter Grundlagenforschung kann die Untersuchung vom Untersuchungsleiter auf einem zentralen Internetserver bereitgestellt wer-

den, die Administration der Untersuchung kann per Internetverbindung von einem beliebigen Arbeitsplatz aus erfolgen. Die Untersuchungsteilnehmer greifen ebenfalls per Internet auf die Untersuchung zu und beantworten die ihnen dargebotenen Aufgaben. Im Falle experimenteller Untersuchungen wird die Auswahl der dargebotenen Aufgaben vom Server gesteuert, z.B. auch eine zufällige Zuordnung von Teilnehmern zu verschiedenen experimentellen Untersuchungsbedingungen. Die Speicherung der Daten erfolgt auf dem zentralen Server. Falls eine unmittelbare Rückmeldung individueller Ergebnisse vorgesehen ist, erfolgt die dafür nötige Auswertung ebenfalls auf dem Server.

Nutzerprofile

- **Untersuchungsteilnehmer:** Untersuchungsteilnehmer müssen lediglich die Möglichkeit haben, die dargebotenen Aufgaben zu beantworten. Es ist in diesem Szenario denkbar, dass Teilnehmer sich nicht selbst anmelden müssen, sondern Nutzer-IDs automatisch vom System vergeben werden. Die Vergabe eines Nutzernamens mit Passwort kann eventuell sinnvoll sein, um einen Zugriff auf eine individuelle Ergebnisrückmeldung zu ermöglichen. Zusätzlich zum Zugriff auf die Untersuchung selbst und eine individuelle Rückmeldung kann Teilnehmern über eine authentifizierte Anmeldung auch Zugriff auf zusammenfassende Ergebnisse der gesamten Untersuchung gegeben werden.
- **Systemadministrator (Untersuchungsleiter):** Beim Untersuchungsleiter liegen die vollen Zugriffsrechte auf das System, die zur Einrichtung der Untersuchung, der Verwaltung des Untersuchungsmaterials und zum Abruf und der Weiterverarbeitung der erfassten Daten notwendig sind.

Testsituation

Die Testdurchführung bzw. Untersuchungsteilnahme erfolgt in diesem Szenario in Einzelsitzungen an den jeweiligen Computern, von denen aus Teilnehmer auf die Untersuchung zugreifen – z.B. von zu Hause oder vom Arbeitsplatz. In der Testsituation gibt es somit keine externe Kontrolle einer instruktionsgemäßen Aufgabenbearbeitung. Um mögliche unerwünschte Folgen dieser mangelnden Kontrolle zumindest teilweise zu kontrollieren und bei der Auswertung zu berücksichtigen, können Variablen wie Bearbeitungszeiten, IP-Adresse und Browser des Nutzers vom Server mit aufgezeichnet werden. Diese können z.B. Hinweise auf längere Unterbrechungen der Bearbeitung oder auf mehrfache Teilnahme von demselben Computer aus liefern.

7 Technische Lösungen für ein computer- und internetbasiertes Assessment-System

Jean-Paul Reeff

In diesem Kapitel werden die bisher beschriebenen Anforderungen an eine technologiebasierte[5] Kompetenzdiagnostik zusammengefasst und Konsequenzen für die Architektur einer entsprechenden Plattform abgeleitet. Einige Anforderungen ergeben sich eher implizit aus den einführenden Kapiteln 1 bis 4; explizit formuliert sind weitere Anforderungen in Kapitel 5 und konkretisiert in den in Kapitel 6 beschriebenen Anwendungsszenarien. Zusätzlich werden in diesem Kapitel einige technologienahe Anforderungen beschrieben sowie die Koppelung eines technologiebasierten Assessmentsystems an andere relevante Systeme skizziert. Ziel ist die funktionale Analyse aller relevanten Etappen computergestützter Kompetenzmessung und ihre Abbildung in einer adäquaten Systemarchitektur.

Bereits in den Kapitel 1 und 2 wird eine wesentliche Vorentscheidung für die mögliche Architektur einer technologiebasierten Kompetenzdiagnostik vorbereitet. Die Vielzahl der dort beschriebenen Anlässe, Kontexte und Zielsetzungen für die Erfassung von Kompetenzen, und – damit verbunden – die Unterschiedlichkeit der Akteure, die zu unterschiedlichen Zeitpunkten in Teilaspekte der Kompetenzerfassung einbezogen sind, bedingt unmittelbar eine entsprechende Vielzahl an zum Teil domänenspezifischen Funktionalitäten auf Systemseite. Dieser Herausforderung kann man auf zwei prinzipiell unterschiedliche Arten begegnen: Entweder entwickelt man mehrere unabhängige Systeme, die auf den jeweiligen Kontext und das Ziel der Kompetenzerfassung maßgeschneidert werden, oder man entwickelt eine generische Plattform, die möglichst alle gemeinsamen „Bausteine" von Kompetenzdiagnostik umfasst und eine einfache „Parametrisierung" für den gewünschten Einsatzfall erlaubt. Ausgehend von den grundlegenden Überlegungen zur technologiebasierten Kompetenzdiagnostik in den Kapiteln 3 und 4 sowie aufgrund ihrer Analyse der derzeitigen Situation in Deutschland empfehlen Hartig, Kroehne & Jurecka (Kap. 5, S. 59) „die Entwicklung eines Basissystems für computer- und netzwerkbasiertes Assessment". Ziel ist dabei „die Entwicklung einer grundlegenden technischen Basis, auf der konkrete Instrumente für spezifische Fragestellungen aufbauen können, ohne dass die grundlegendsten Entwicklungsschritte jedes Mal neu durchlaufen werden müssen". Diese Forderung gilt als grundlegende Anforderung für die Darstellungen in diesem Kapitel.

---

[5] Der Terminus „technologiebasiert" wird in diesem Kapitel in einer allgemeinen Form verwendet und umfasst u.a. die bisher verwendeten Termini „computerbasiert" resp. „internetbasiert". Er ermöglicht aber auch die Beschreibung anderer Technologien (Handhelds, electronic ink, usw.).

7.1 Anforderungen an ein technologiebasiertes Basissystem zur Kompetenzerfassung

Einer der systematischsten Versuche, Assessmentsysteme zu beschreiben, stammt von Allmond, Steinberg & Mislevy (2002). Ihre Analyse setzt auf dem vom *Educational Testing Service* (ETS) entwickelten *Evidence-Centered assessment Design* (ECD) Framework auf (Steinberg et al., 2000) und mündet in einer "Vier-Prozess-Architektur". Nach Allmond et al. (2002) muss jedes Assessmentsystem – zumindest in minimaler Ausprägung – folgende Prozesse abbilden:

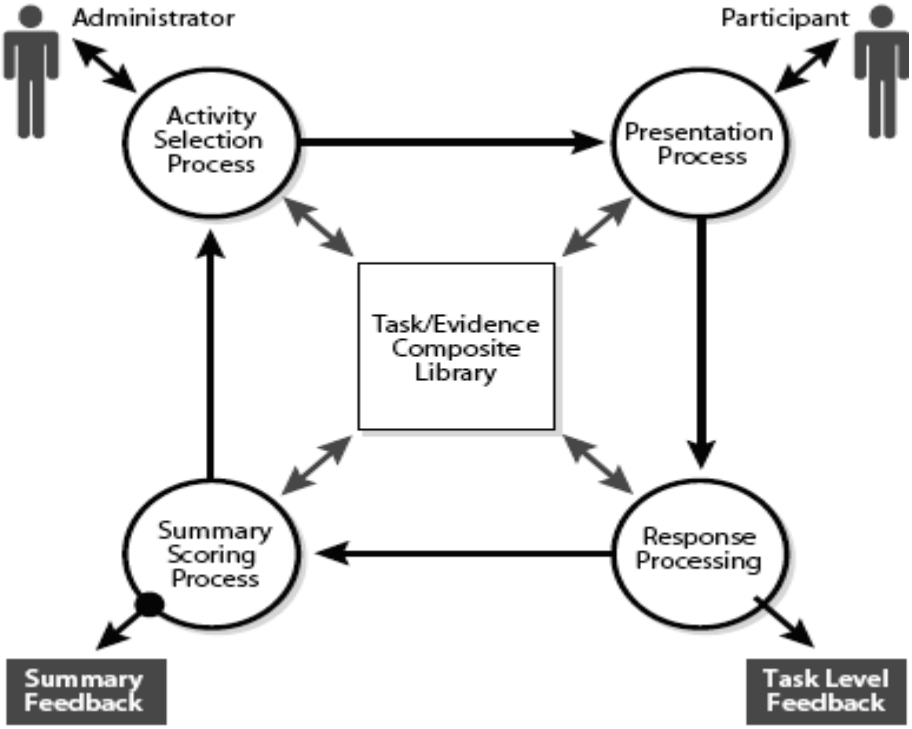
(1) Activity Selection Process
(2) Presentation Process
(3) Response Processing
(4) Summary Scoring Process.

Der *Activity Selection Process* umfasst die Auswahl und Sequenzierung von Aufgaben aus der *Task/Evidence Composite Library* (eine Datenbank, die Aufgaben/Items selbst sowie alle relevanten Metadaten zu den Aufgaben enthält). Der *Presentation Process* steuert den gesamten Prozess der Aufgaben-/Test*darstellung* sowie die Interaktion des Teilnehmers mit dem System. *Response Processing* ist zuständig für eine erste Aufbereitung der Teilnehmerdaten (Identifikation und erste Beurteilung von responses). Die so aufbereiteten Daten werden an den *Summary Scoring Process* zur weiteren Auswertung übergeben. Die auf einen ersten Blick künstlich wirkende Aufspaltung der beiden letztgenannten Prozesse ist bei genauerer Betrachtung essentiell für die Mehrfachnutzung (reusability) von Aufgaben in unterschiedlichen Kontexten.

Nach Almond et al. entsteht ein vollständiger Assessment-Zyklus durch die Interaktion dieser vier Prozesse und unter Kontrolle von mindestens zwei Akteuren, dem *Administrator* und dem *Teilnehmer*. Der Administrator ist verantwortlich für Design und Durchführung des Assessment, sowie für alle Fragen von Konfiguration und Parametrisierung der Testdurchführung. Teilnehmer ist die Person, deren Kompetenzen erfasst werden sollen.

Das skizzierte Modell dient in diesem Kapitel als erste Präzisierung des in Kapitel 5 geforderten Basissystems und erlaubt eine Strukturierung der dort beschriebenen, einzelnen Anforderungen. Der „Activity Selection Process" wählt eine Aufgabe aus (Items, Itemzusammenstellungen oder andere Aktivitäten) und leitet sie zur Aufgabendarstellung an den „Presentation Process" weiter. Nachdem der Teilnehmer die Bearbeitung der Aufgabe abgeschlossen hat, sendet der „Presentation Process" das Ergebnis („Work Product") zum „Response Processing". Dieser Prozess identifiziert die essentiellen Charakteristika der Ergebnisse und leitet diese zum „Summary Scoring Process" weiter. Dieser aktualisiert die bisher aufgezeichneten Ergebnisse, anhand derer Annahmen über die Kompetenzen des Probanden angestellt werden. Der „Activity Selection Process" entscheidet dann, basierend auf den aktuellen Annahmen über den Probanden oder anderen Kriterien, über den weiteren Verlauf (Abbildung 7.1 skizziert die Zusammenhänge).

Abbildung 7.1: Die vier grundlegenden Prozesse im Assessment-Zyklus (aus: Almond, Steinberg & Mislevy; 2002).



Im Lichte dieses Modells kann man die spezifischeren Anforderungen aus den vorigen Kapiteln folgendermaßen rekapitulieren: Der in Kapitel 5 beklagten Situation einer wenig professionellen und unsystematischen Softwareentwicklung kann man in einem ersten Ansatz mit dem Entwurf einer modular aufgebauten Plattform (eines „Basissystems") begegnen, die sich an einer konzeptuellen Architektur wie ECD oder "Vier-Prozess-Architektur" orientiert. Die aus einer generischen Anforderungsanalyse entstandenen Spezifikationen für Assessmentsysteme sollten in einer konkreten Plattform transparent und modular abgebildet sein. Dies heißt nicht notwendigerweise, dass jedem Prozess ein einziges Softwaremodul zugeordnet wird; eine modulare Implementierung sollte aber alle relevanten Prozessaspekte abbilden. Die Plattform sollte auch in der Lage sein, mit solchen Spezifikationen verbundene und sich dynamisch weiterentwickelnde Standards schnell und transparent integrieren zu können.

Um eine solche Plattform möglichst generell einsetzen zu können, sollte sie betriebssystemunabhängig sein und möglichst standardisierte Schnittstellen zu anderen Softwaretools und betriebssystemübergreifenden Plattformen haben (z.B. e-Learning-Plattformen, s.u.). Spezifischere Anforderungen im Rahmen des skizzierten Modells können wie folgt geordnet werden:

Ad 1) Activity Selection Process
Die geforderte hohe Flexibiltät auf der Ebene von Datenformaten und Datenschnittstellen muss sich insbesondere in der Implementierung des Activity Selection Process widerspiegeln. Eine benutzerfreundliche Auswahl und Sequenzierung von Aufgaben setzt eine transparente Organisation und Beschreibung der Items/Aufgaben voraus. Dies führt im Falle von TBA zu der zusätzlichen Forderung von leistungsfähigen Autorensystemen, die in der Lage sind, auch komplexere Typen („templates") von Items/Aufgaben zu generieren und es so auch Nicht-Informatikern ermöglicht entsprechende Aufgaben selbst zu generieren. Dieser Aspekt des Authoring ist in der "Vier-Prozess-Architektur" nicht ausreichend berücksichtigt, hat aber im Rahmen einer technologiebasierten Kompetenzerfassung zentrale Bedeutung. Das Autorensystem muss so konstruiert sein, dass es nicht nur die Aufgabentypen selbst und dazu korrespondierende Aufgaben generieren kann, sondern auch automatisch und/oder benutzergesteuert alle relevanten Metadaten ablegen kann.

Die in Kapitel 5 beschriebene Berücksichtigung von Aufgabenhierarchien sollte in der Implementierung des Activity Selection Process gelöst werden. Im Activity Selection Process, wie in allen anderen Prozessen auch, muss ein an Funktionen und Benutzertypen orientiertes Berechtigungssystem implementiert sein.

Ad 2) Presentation Process
Die in Kapitel 5 geforderte „automatische Anmeldung" von Benutzern, die in externen, bestehenden Systemen verwaltet werden, könnte im Berechtigungssystem des Presentation Process gelöst werden. Für größere Teilnehmergruppen und komplexere Testdesigns empfielt sich die Implementierung eines separaten Moduls zur Teilnehmerverwaltung.

In einer Offline-Testdurchführung ist die „Aufgabenentnahme" aus der „Task/Evidence Composite Libray" ein relativ trivialer Prozess. Im Hinblick auf zukünftige internetbasierte Testdurchführungen, mit Zugriff auf verteilte Itembanken unterschiedlicher Eigentümer, sollte der Presentation Process leistungsfähige Mechanismen zur Auswahl von Aufgaben und Items beinhalten.

Es ist offenkundig, dass auf der Ebene des Presentation Process auch wesentliche Aspekte von Datenschutz berücksichtigt werden müssen, sowohl was die Teilnehmer, als auch was die Aufgaben bzw. Items anbelangt.

Ad 3) Response Processing
Im Response Process äußert sich ein fundamentaler Unterschied zwischen Kompetenzerfassung in Large-Scale-Surveys einerseits und in einem Kontext von Individualdiagnostik andererseits. In letzterem Fall wird es häufig notwendig oder wünschenswert sein, Zwischenresultate zur Steuerung des Testablaufs (adaptives Testen) zeitnah zu verwenden.

Ebenso muss die in Kapitel 5 geforderte Anpassung an Lehr-/Lernkontexte und die damit häufig verbundene zeitnahe Bereitstellung von Testresultaten in z.T. ebenfalls computerbasierten Lernumgebungen hier implementiert werden. Das für die Abbildung des Response Process zuständige Modul muss die dafür

notwendigen Mechanismen und Schnittstellen zum Presentation process sowie zu evt. anderen Systemen vorsehen.

Ad 4) Summary Scoring Process
Auf der Ebene des Summary Scoring Process muss der Forderung unterschiedlicher Auswertungsmodelle genügt werden. Ebenso müssen Schnittstellen zu den wichtigsten statistischen Auswertungspaketen bereitgestellt werden.

Die ebenfalls in Kapitel 5 erhobene Forderung, auch in einem technologiebasierten System zur Kompetenzerfassung weiterhin den Einbezug menschlicher Beurteiler zu ermöglichen, lässt sich wahrscheinlich am besten auf der Ebene des Summary Scoring Process lösen, indem hier eine Schnittstelle zu entsprechenden Systemen implementiert wird. So hat beispielsweise das IEA Data Processing Center in Hamburg bereits weitgehende Vorarbeiten für eine computergestützte Auswertung (eingescannter) offener Antworten geleistet. Die dort entwickelte Software könnte somit ein konkreter Baustein bei der Implementierung des Summary Scoring Process sein.

Schließlich sollten insbesondere im Zusammenhang mit Schulrückmeldungen im Summary Scoring Process eines technologiebasierten Systems leistungsfähige Tools für ein benutzerfreundliches Reporting zur Verfügung gestellt werden.

7.2 TAO: Ein Konzept und eine Plattform für technologiebasierte Kompetenzmessung

Sowohl das ECD Framework als auch die "Vier-Prozess-Architektur" reflektieren in erster Linie eine systematische und ingenieursmässige Vorgehensweise von ETS beim Design von Assessments. In beiden Konzepten wird technologiebasierte Kompetenzerfassung zwar mitbedacht, die Vorbereitung und Konzeptualisierung einer entsprechenden Plattform war aber nicht die vorrangige Zielsetzung dieser allgemeineren Anstrengung. Der systematische Entwurf einer an ECD und „Vier-Prozess-Architektur" orientierten Plattform steht jedenfalls noch aus. Es darf zwar davon ausgegangen werden, dass die ETS-internen Entwicklungsarbeiten sich am allgemeineren Modell orientieren, die entsprechenden Arbeiten sind aber bei weitem nicht abgeschlossen, eine öffentliche Nutzung der Plattform ist nicht gegeben.

Eine Analyse unterschiedlicher existierender Plattformen im Hinblick auf Übereinstimmung mit dem ECD Framework und der "Vier-Prozess-Architektur" ist aufgrund der mangelnden Dokumentation bzw. öffentlichen Verfügbarkeit der dahinter liegenden Modelle nur partiell möglich. Die meisten komplexeren Plattformen sind aber auch dediziert auf bestimmte (manchmal umfangreiche Klassen von) Anwendungen hin entwickelt worden. Ein wichtiges Beispiel sind in diesem Zusammenhang die von der University of Auckland entwickelten Assessment Tools for Teaching and Learning (asTTle, *www.tki.org.nz/r/asttle*). Die Plattform wurde im Auftrag des neuseeländischen Bildungsministeriums entwickelt und zielt darauf, ein systematisches Assessment von Literacy und Numeracy bei

4 bis 12jährigen Schülern zu ermöglichen. Obwohl die Plattform deutlich über alles hinausgeht, was zurzeit in Deutschland verfügbar ist, fällt es schwer zu erkennen, wie eine Nutzung/Erweiterung der Plattform in anderen Kontexten möglich wäre. Neben der technologischen Unwägbarkeit erschweren außerdem Fragen des Urheberrechts sowie schwer zu quantifizierende Finanzierungen für notwendige Erweiterungen eine Beurteilung der Plattform als Ausgangspunkt für das in Kapitel 5 skizzierte Basissytem.

Eine ähnliche Situation findet man, wenn man sich an den Tools existierender – nationaler oder internationaler – large-scale assessments orientiert. Weder die in PISA 2006 genutzte Plattform für computer-based assessment of scientific literacy (CBAS), noch Plattformen, wie sie in einigen US-Bundesstaaten für nationale Monitoringprojekte eingesetzt werden (z.B. in Virginia, Kentucky, u.a.), sind offen oder ausbaufähig genug, um als möglicher Ausgangspunkt für das geforderte Basissystem zu dienen.

Die einzige dem Autor bekannte, systematisch entwickelte und offen dokumentierte generische Architektur, die zudem zu einer erfolgreichen Implementierung führte, ist das in Luxemburg entwickelte TAO-System (Martin, Latour, Burton, Busana & Vandenabeele, 2005; Plichart, Jadoul, Vandenabeele & Latour, 2004). Die gemeinsam von Assessmentexperten und Software-Ingenieuren entwickelte Plattform befindet sich im Stadium eines ausführlich getesteten Forschungsprototyps. Die Plattform ist *open-source*, d.h. der gesamte Quellcode wird allen potentiellen Anwendern frei zur Verfügung stehen, sofern sie die gewählte Open-Source-Lizenz akzeptieren.

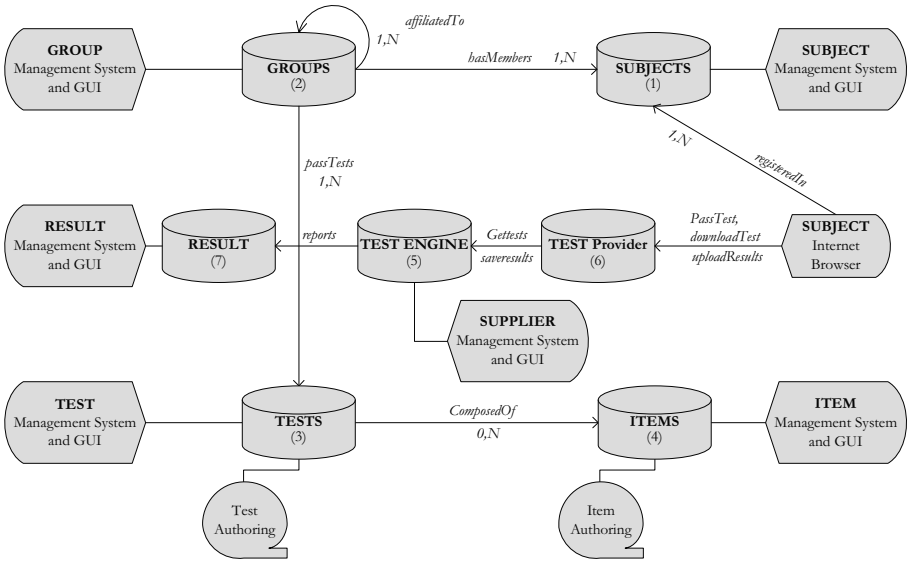
Obwohl die TAO-Verantwortlichen weder auf das ECD Framework, noch auf die „Vier-Prozess-Architektur" zurückgreifen, kommen sie nach umfangreichen Analysen und Befragungen unterschiedlicher Nutzergruppen zu einem sehr ähnlichen Ergebnis. Im Unterschied zu den ETS-Anstrengungen, die auf generische Charakteristika beliebiger Assessments abzielten, stand bei TAO die konkrete Entwicklung einer technologiebasierten Plattform für internetbasiertes Testen von Anfang an im Vordergrund. Dies führt zu einer deutlich höheren Auflösung in den Analysen und zu einer detaillierteren (und software-näheren!) Architektur, als dies beispielsweise in der „Vier-Prozess-Architektur" der Fall ist.

Die Ambition, eine generelle und sehr generische Plattform zu entwerfen und zu implementieren, hat aber auch zu einer anderen Vorgehensweise beim Design der Architektur geführt, als dies bei vielen anderen Plattformen der Fall war. Da bei TAO keine dringenden kurz- oder mittelfristigen Anwendungen berücksichtigt werden mussten (im Gegensatz etwa zu PISA/CBAS oder den oben erwähnten state-wide assessments in einigen US-Bundesstaaten), konnten die Verantwortlichen einen sehr systematischen Top-Down-Analyse- und -Designansatz verfolgen. Dieser Ansatz hat zu einer Grundarchitektur von TAO geführt, die fünf wesentliche Aspekte beim Assessment unterscheidet. Diese Aspekte haben sich bei der Implementierung der Plattform in fünf verschiedenen Modulen niedergeschlagen:

(1) einem *Subjects*-Modul, das alle Benutzerinformationen verwaltet;
(2) einem *Group* (of subjects)-Modul, in dem Testteilnehmer nach unterschiedlichen Kriterien organisiert werden können;
(3) einem *Test*-Modul, das Items nach unterschiedlich möglichen Testdesigns zu ausführbaren Tests zusammenführt;
(4) einem *Item*-Modul, das das Generieren von unterschiedlichen Itemarten und dazugehörigen Items ermöglicht;
(5) einem *Results*-Modul, das für die Auswertung und Präsentation von Resultaten zuständig ist.

Zusätzlich übernimmt die *Test Engine* die Koordination zwischen den Modulen und verwaltet alle relevanten Daten und Metadaten, die für eine erfolgreiche Durchführung des Tests notwendig sind. Abbildung 7.2 gibt einen Überblick.

Abbildung 7.2: Schematische Übersicht der TAO Plattform Architektur (aus: Plichart, Jadoul, Vandenabeele & Latour, 2004).



Die im Vergleich zur „Vier-Prozess-Architektur" komplexere Darstellung hängt wesentlich damit zusammen, dass in der Analyse- und Designphase Implementierungsaspekte systematisch mitbedacht wurden. Eine Abbildung der wesentlichen Komponenten der „Vier-Prozess-Architektur" auf die TAO-Architektur ist allerdings vergleichsweise einfach:

- Die mit dem Activity Selection Process verbundene Auswahl und Sequenzierung von Items wird in TAO vom Testmodul übernommen. Der Presentation Process ist Bestandteil des Item-Moduls (Item Management System and GUI) und wird zur Testdurchführungszeit von der Test Engine koordiniert. Response Processing ist in TAO genau wie der Presentation Process auf das Item-Modul und das Test-Modul verteilt, beide Prozesse bleiben a-

ber im Sinn der "Vier-Prozess-Architektur" auch in TAO eigenständige Prozesse. Der Summary Scoring Process ist eins zu eins dem Results-Modul zuzuordnen.

- TAO sieht umfangreiche Mechanismen zum "Abonnieren" von Items aus unterschiedlichen Datenbanken vor. Die gesamte so generierte Itembank, inklusive Metadaten, entspricht der Task/Evidence Composite Library in der "Vier-Prozess-Architektur". Im allgemeinen Fall wird sie in TAO eine hochkomplexe verteilte Datenbank sein, im einfachsten Fall wird sie sich auf die mit dem Itemmodul assoziierte Datenbank reduzieren lassen.
- Über die "Vier-Prozess-Architektur" hinaus stellt TAO spezialisierte Funktionen für das Generieren von Items zur Verfügung (Item Authoring Tool). Außerdem stellen die Module *Subjects* und *Groups* leistungsfähige Mechanismen zur Verwaltung der Teilnehmer zur Verfügung. Die separate Verwaltung von Teilnehmerdaten in einem unabhängigen Modul (Subjects) legt darüber hinaus den Grundstein für eine dem Datenschutz genügende Architektur.

Eine detaillierte Darstellung der Gesamtfunktionalität von TAO geht weit über dieses Kapitel hinaus (vgl. hierzu Plichart et al., 2004 und Martin et al., 2005). Insgesamt lässt sich aber sagen, dass alle wesentlichen Anforderungen an ein Assessment-System gemäß ECD und "Vier-Prozess-Architektur" sowohl von der konzeptuellen TAO-Architektur als auch von der zurzeit vorliegenden Implementierung erfüllt werden. Darüber hinaus hat TAO mehrere Aspekte konkreter und detaillierter realisiert. Über die generischen Anforderungen hinaus verfügt TAO über Mechanismen zur Erfüllung aller in diesem Abschnitt erwähnten spezifischen Anforderungen.

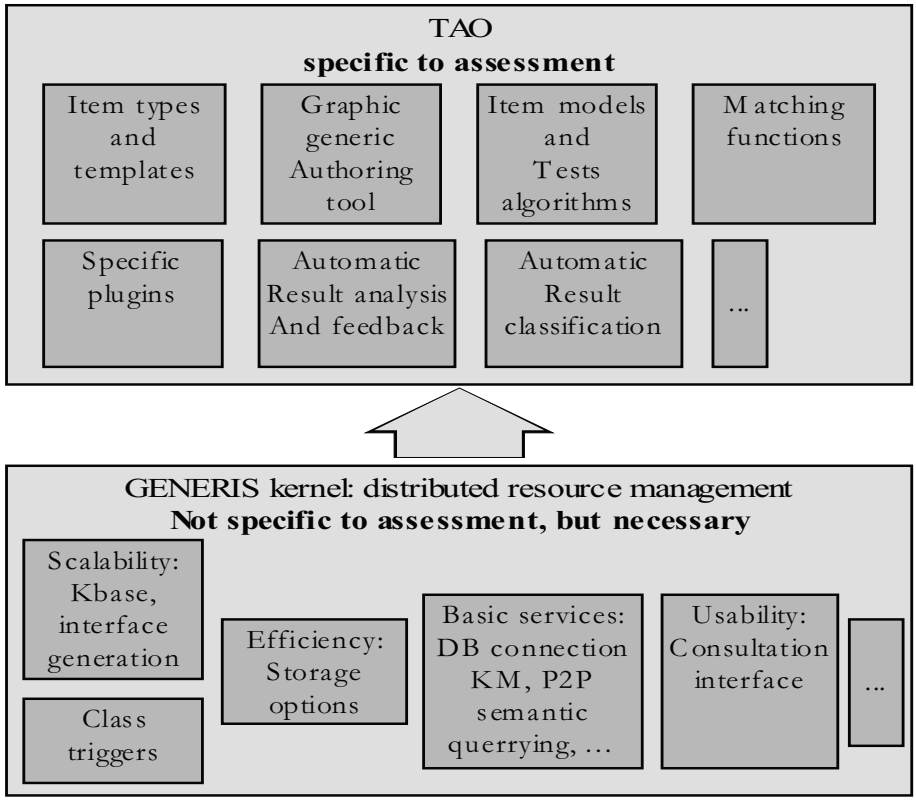
Einige weitere Anforderungen, die insbesondere im Kapitel 5 beschrieben sind, seien zusätzlich kurz beschrieben:

- TAO kann sowohl offline wie online betrieben werden. Im Online-Betrieb werden alle Varianten von lokalen (Kleinst-)Serverlösungen bis hin zu internetbasierten Testdurchführungen mit Zugriff auf komplexe verteilte Datenbanksysteme unterstützt.
- Eine kollaborative Entwicklung von Aufgaben wird unterstützt. Ein in Luxemburg beantragtes Forschungsprojekt zielt sogar auf eine systematische Softwareunterstützung des gesamten Prozesses zur Entwicklung von Bildungsstandards, inklusive der kollaborativen Entwicklung von dazugehörigen Messinstrumenten.

Aus software-technischer Sicht sei noch auf die Zwei-Layer-Implementierung von TAO hingewiesen. Ein unterer Layer, der so genannte Kernel, ist nicht spezifisch für Assessment-Anwendungen, sondern ist ein anwendungsunspezifischer Bestandteil von TAO, der grundlegende Funktionen und Services für das Verwalten verteilter Ressourcen zur Verfügung stellt. Über den ästhetischen Wert einer solchen Lösung hinaus bietet der Ansatz gerade im Rahmen einer Open-Source-Strategie den Vorteil, dass langfristig eine hohe Qualität der Weiternetwicklung und Wartung des Kernels auch von anderen Communities als der Assessment-Community betrieben werden kann.

Über dem Kernel liegen dann – auf einer zweiten Ebene – die TAO-Komponenten, die assessmentspezifische Funktionen haben. Abbildung 7.3 zeigt den Aufbau und ausgewählte Funktionen auf beiden Ebenen.

Abbildung 7.3: Aufbau und ausgewählte Funktionen des Kernels und der TAO-Komponenten.



7.3 Stärken und Schwächen von TAO

Aufgrund einer systematischen Top-Down-Analyse und einer damit verbundenen, erfolgreichen Erstimplementierung zeigt TAO keine grundlegenden Schwächen. Die Breite des Konzepts und die daraus resultierenden vielfältigen Einsatzmöglichkeiten sind die herausragende Stärke von TAO. Das systematische, an internationalen Standards orientierte Software-Engineering ist als weiteres positives Merkmal hervorzuheben. Schließlich ist aus Anwendersicht die Entscheidung, TAO als Open-Source-Produkt zu entwickeln, sicherlich nur zu begrüßen. Während TAO also in seinen Grundzügen überzeugt, bleiben noch einige Herausforderungen zu bewältigen, bzw. Schwachstellen zu beseitigen:

- Obwohl TAO bereits in größerem Umfang getestet worden ist (zuletzt in einem Online-Test mit über 4.000 Schülern, bis zu 1.000 davon gleichzeitig),

haben die bisherigen Tests einige Instabilitäten und Unzulänglichkeiten aufgedeckt. Einige konnten in der Phase des Prototyping beseitigt werden, insgesamt scheint es aber wünschenswert und notwendig, den TAO-Kernel im Hinblick auf die gemachten Erfahrungen zu reengineeren. Eine Beschreibung dieses Reengineering sprengt den Rahmen dieses Textes; insgesamt kann man aber sagen, dass sich die gewählte Architektur und Funktionalität weitestgehend bewährt hat, aber auf einer softwaretechnischen Ebene Modifikationen vorgenommen werden müssen bzw. zum Teil auch andere Technologien zum Einsatz kommen sollten.

- Auf der assessmentspezifischen Ebene ist die Bereitstellung eines leistungsfähigen Autorensystems ein wichtiges Desiderat. Die Bemühungen der Entwickler gehen zurzeit in die Richtung eines generischen Autorensystems, das es ermöglichen soll, Templates für beliebige Aufgabentypen über das Autorensystem zu generieren.
- In diesem Zusammenhang muss auch geklärt werden, wie gerade für das Authoring die Koppelung mit existierenden, komplexen Softwarepaketen gewährleistet werden kann. Zur Illustration: Ein Versuch der Koppelung mit AgentSheets zum Zweck der Entwicklung von komplexen Problemlöseaufgaben zeigte, dass dies zwar sehr wohl möglich ist, die Palette an Koppelungsmechanismen aber noch erweitert werden muss[6].
- In ähnlicher Weise empfiehlt es sich zu prüfen, ob ein softwaregestütztes Scoring von Antworten durch menschliche Scorer (z.B. bei offenen Fragen in Lesetests) verstärkt innerhalb von TAO unterstützt werden soll oder ob man das Problem eher durch Koppelung mit existierenden Plattformen (wie beispielsweise vom IEA-DPC entwickelt, s.o.) lösen sollte.
- Die Sicherheitsmechanismen bezüglich Itemsicherheit müssen für allerhöchste Schutzbedingungen (high-stakes-testing, kommerzielle Anwendungen, u.a.) noch weiter verbessert werden. Dies gilt auch für die Einbindung von Algorithmen, die ggf. nicht open-source sind, wie das für eine Zusammenarbeit bspw. mit kommerziellen Testanbietern notwendig ist.
- Auch im Lichte der in Kapitel 5 geforderten Barrierefreiheit muss nachgearbeitet werden. Um hier zu sinnvollen Lösungen zu kommen, ist es erforderliche, die Forderung genauer zu spezifizieren und auf dieser Basis Lösungen zu implementieren, die sich an international entwickelnden Standards orientieren[7].
- Schließlich sollte im Hinblick auf aktuelle Entwicklungen geprüft werden, wie sich TAO in allgemeinere E-Learning-Systeme oder Plattformen für kollaboratives Arbeiten integrieren lässt[8].

---

[6] Siehe z.B.hierzu beispielsweise http://www.cs.colorado.edu/~ralex/Portfolio.pdf und http://www. agentsheets.com/).

[7] Siehe beispielsweise http://www.w3.org/TR/1999/WAI-WEBCONTENT-19990505/ oder BITV).

[8] Siehe z.B. CROQUET, http://www.opencroquet.org/index.html.

7.4 Zusammenfassung und Schlussfolgerung

Ausgehend von der generischen Anforderungsanalyse des Evidence Centered assessment Design (ECD) resp. der „Vier-Prozess-Architektur" wurden die aus den vorherigen Kapitel ableitbaren Anforderungen an ein technologiebasiertes Basissystem zur Kompetenzerfassung in einer softwarenäheren Form dargestellt. Mit der anschließend beschriebenen TAO-Plattform steht ein solches Basissystem in Form eines gut erprobten Forschungsprototyps zur Verfügung. TAO erfüllt alle grundlegenden Anforderungen, der TAO-Kernel muss aber hinsichtlich Systemstabilität und Effizienz reengineert werden. Die Bereitstellung eines leistungsfähigen Autorensystems so wie die Verbesserung einiger technologischer Aspekte im assessmentspezifischen Teil der Plattform sind weitere Desiderata. Unter Berücksichtigung der vorzunehmenden Änderungen stellt sich TAO als eine äußerst leistungsfähige Implementierung des in Kapitel 5 geforderten Basissystems dar. Die Tatsache, dass die Plattform open-source ist, macht sie in verstärktem Maße attraktiv für das breite Spektrum an Szenarien zur technologie-basierten Kompetenzerfassung, wie sie in dieser Expertise diskutiert werden.

Literatur

Ackermann, T. A., Evans, J., Park, K.-S., Tamassia, C. & Turner, R. (1999). Computer assessment using visual stimuli: A test of dermatological skin disorders. In: F. Drasgow & J. B. Olson-Buchanan (Eds.), Innovations in computerized assessment (S. 137-150). Mahwah, NJ: Lawrence Erlbaum.

Ackerman, T. A., Gierl, M. J. & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. Educational Measurement: Issues and Practice, 22, S. 37-53.

Adams, R. (Ed.) (2005). PISA 2003 Technical Report. Paris: OECD.

Adams, R. & Wu, M. (2002). (Eds.). PISA 2000 technical report. Paris: OECD.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. Journal of Technology, Learning, and Assessment, 1(5) (http://www.jtla.org).

Amelang, M. & Schmidt-Atzert, L. (2006). Psychologische Diagnostik und Intervention (4. Auflage). Berlin: Springer.

American Psychological Association (1986). Guidelines for computer-based tests and interpretations. Washington, DC: APA.

Baker, E. & O'Neil, H. F. (2002). Measuring problem solving in computer environments: Current and future states. Computers in Human Behavior, 18, S. 609-622.

Barak, A. & English, N. (2002). Prospects and limitations of psychological testing on the Internet. Journal of Technology in Human Services, 19, S. 65-89.

Barret, G. V. & Depinet, R. L. (1991). A reconsideration of testing for competence rather than for intelligence. American Psychologist, 46, S. 1012-1024.

Baumert, J., Stanat, P. & Demmrich A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In: Deutsches PISA-Konsortium (Hrsg.) (2001), PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich (S. 11-38) Opladen: Leske & Budrich.

Beaton, E. & Allen, N. (1992). Interpreting scales through scale anchoring. Journal of Educational Statistics, 17, S. 191-204.

Beck, B. & Klieme, E. (Hrsg.) (2007). Sprachliche Kompetenzen – Konzepte und Messung. Weinheim: Beltz

Bennett, T. (2005). The media sensorium: cultural technologies, the senses and society. In: M. Gillespie (Ed.), Understanding Audiences. Milton Keynes: The Open University Press.

Bennett, R. E., Goodman, M., Hessinger, J., Kahn, H., Ligget, J., Marshall, G. & Zack, J. (1999). Using multimedia in large-scale computer-based testing programs. Computers in Human Behavior, 15, S. 283-294.

Birnbaum, H. (2000) (Hrsg.). Psychological experiments on the Internet. San Diego, CA: Academic Press.

Blömeke, S. (2003). Medienpsychologische Kompetenz. Theoretische Grundlagen und erste empirische Befunde. Empirische Pädagogik, 17, S. 196-216.

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F. et al. (2004). Methematische Kompetenz. In: PISA-Kosortium Deutschland (Hrsg.), PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs (S. 47-92). Münster: Waxmann.

Bodman, S. M. & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. Journal of Educational Computing Research, 31, S. 51-60.

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004): The Concept of Validity. Psychological Review, 111, S. 1061-1071.

Brehmer, B. & Dörner, D. (1993). Experiments With Computer-Simulated Microworlds: Escaping Both the Narrow Straits of the Laboratory and the Deep Blue Sea of the Field Study. Computers in Human Behavior, 9, S. 171-184.

Breland, H. M. & Lytle, E. G. (1990, April). Computer-assisted writing skill assessment using WordMap. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Breuer, K. & Satish, U. (2003). Emergency Management Simulations – An approach to the assessment of decision making processes in complex dynamic crisis environments. In: J. J. Gonzaléz (Ed.), From modeling to managing security – A system dynamics approach (S. 145-156). Kristiansand: Norwegian Academic Press.

Buchanan, T. & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World-Wide Web. British Journal of Psychology, 90, S. 125-144.

Burstein, J., Kaplan, R., Wolff, S. & Lu, C. (1997). Automatic scoring of advanced placement biology essays. Princeton, NJ: ETS.

Cattell, R. B. & Warburton, F. W. (1967). Objective personality and motivation tests. Urbana: University of Illinois Press.

Chung, G. K. W. K., Baker, E., Brill, D. G., Sinha, R. & Saadat, F. (2003, November). Automated assessment of domain knowledge with online knowledge mapping. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL.

CITO (2005). Cito-Sprachtest Zur Sprachstandsfeststellung bei Kindern im Vorschulalter. Arnhem: CITO (http://www.cito.nl/de_info/Sprachtest-Prospekt-06-2005.pdf).

Clyman, S. G., Melnick, D. E. & Clauser, B. E. (1995). Computer-based case simulations. In: E. L. Mancall & P. G. Bashook (Eds.), Assessing clinical reasoning: The oral examination and alternative methods (S. 139-149). Evanston, IL: American Board of Medical Specialities.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, S. 281-302.

Csapó, B. (2004). Knowledge and competencies. In: J. Letschert (Ed.), The integrated person. How curriculum development relates to new competencies (S. 35-49). Enschede: CIDREE/SLO.

Donovan, M. A, Drasgow, F. & Probst, T. M. (2000). Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. Journal of Applied Psychology, 85, S. 305-313.

Dörner, D., Kreuzig, H. W., Reither, F. & Stäudel, T. (1983). Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität. Bern: Huber.

Dörner, D. & Preußler, W. (1990). Die Kontrolle eines einfachen ökologischen Systems. Sprache & Kognition, 9, S. 205-217.

Dörner, D., Schaub, H. & Strohschneider, S. (1999). Komplexes Problemlösen – Königsweg der Theoretischen Psychologie? Psychologische Rundschau, 50, S. 198-205.

Dörner, D. & Wearing, A. T. (1995): Complex Problem-Solving: Towards a (computersimulated) Theory. In: J. Funke & P. Frensch (Eds.), Complex Problem Solving. The European Perspective (S. 65-99). Hillsdale, New Jersey: Erlbaum.

Drasgow, F. (2002). The work ahead: A psychometric infrastructure for computerized adaptive tests. In: C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), Computer-based testing. Building the foundation for future assessments (S. 1-35). Mahwah, NJ: Lawrence Erlbaum.

Duden (2001). Fremdwörterbuch. 7., neu bearbeitete und erweiterte Auflage. Mannheim u.a.: Dudenverlag.

Eggen, T. (in Druck). Adaptive Testing and Item Banking. In: J. Hartig, E. Klieme & D. Leutner (Eds.), Assessment of Competencies in Educational Contexts. Göttingen: Hogrefe & Huber Publishers.

Embretson, S. E. (1983). Construct validity: construct representation vs. nomothetic span. Psychological Bulletin, 93, S. 179-197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: application to abstract reasoning. Psychological Methods, 3, S. 380-396.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. American Psychologist, 61, S. 50-55.

Embretson, S. E. & Reise, S. P. (2000). Item response theory for psychologists. Mawah, NJ: Lawrence Erlbaum Associates.

ETS (2005). TOEFL iBT at a glance. Retrieved 26.09.2005, from the World Wide Web (http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_at_a_Glance.pdf).

Fleischer, J., Pallack, A., Wirth, J. & Leutner, D. (2005, September). Vergleichende Schulrückmeldungen im Rahmen der Lernstandserhebungen in Nordrhein-Westfalen. Poster präsentiert auf der 67. Tagung der Arbeitsgruppe Empirisch-Pädagogischer Forschung (AEPF), Salzburg, Österreich.

Flexer, R. W. & Baer, R. M. (2005). Description and Evaluation of a University-Based Transition Endorsement Program. Career Development for Exceptional Individuals, 28, S. 80-91.

Folk, V. G. & Smith, R. L. (2002). Models for delivery of CBTs. In: C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), Computer-based testing. Building the foundation for future assessments (S. 41-66). Mahwah, NJ: Lawrence Erlbaum.

Franzke, M., Kintsch, E. & Kintsch, W. (2005, August). Using summary street to improve reading and writing instruction. Paper presented at the 11th Biennal Conference of the European Association for Research on Learning and Instruction, Nicosia.

Frey, A. (2006). Validitätssteigerungen durch adaptives Testen. Frankfurt am Main: Peter Lang Verlag.

Frey, A., Blunk, H. A. & Banse, R. (2006). PSI-Land: Paarinteraktionsforschung in einer computerbasierten virtuellen Umgebung. Zeitschrift für Sozialpsychologie, 37 (3), S. 151-159.

Frey, A., Carstensen, C. H. & Hartig, J. (2006). BIB-Designs in large scale assessments. Paper presented at the 71th annual meeting of the Psychometric Society in Montréal, June 14-17 2006.

Frey, A., Hartig, J., Ketzel, A., Zinkernagel, A. & Moosbrugger, H. (in Druck). The use of virtual environments based on a modification of the computer game Quake III Arena in psychological experimenting. Computers in Human Behavior.

Fuhrman, S. H. & Elmore, R. F. (Eds.). (2004). Redesigning accountability systems for education. New York: Teachers College Press.

Freudenthaler, H. H., Specht, W. & Paechter, M. (2004). Von der Entwicklung zur Akzeptanz und professionellen Nutzung nationaler Bildungsstandards. Erziehung und Unterricht, 154 (7-8), S. 606-612.

Gaugler, B. B, Rosenthal, D. B, Thornton, G. C & Bentson, C. (1987). Meta-analysis of assessment center validity. Journal of Applied Psychology, 72, S. 493-511.

Gauthier, J.-M. (2005). The Virtual Patient Project. Tisch School of the Arts New York University (http://www.tinkering.net/virtualpatient).

Gershon, R. C. (2005). Computer Adaptive Testing. Journal of Applied Measurement, 6, S. 109-127.

Gijbels, Dochy, Bossche & Segers (2005).

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. Journal of Personality and Social Psychology, 59, S. 1216-1229.

Goldhammer, F. & Hartig, J. (in Druck). Interpretation von Testresultaten und Testeichung. In: H. Moosbrugger & A. Kelava, Test- und Fragebogenkonstruktion. Berlin: Springer.

Gonzalez, E. J. (2003). Scaling the PIRLS Reading Assessment Data. In: M. O. Martin, I. V. S. Mullis & A. M. Kennedy (Eds.): PIRLS 2001 Technical Report (S. 151-168) . Chestnut Hill, MA: Boston College.

Gonzalez, E. J., Galia, J. & Li, I. (2004). Scaling Methods and Procedures for the TIMSS 2003 Mathematics and Science Scales. In: M. O. Martin, I. V. S. Mullis & S. J. Chrostowski (Eds.): TIMSS 2003 Technical Report (S. 252-273). International Association for the Evaluation of Educational Achievement.

Groot, A. S., de Sonneville, M. J. & Stins, J. F. (2004). Familial influences on sustained attention and inhibition in preschoolers. Journal of Child Psychology and Psychiatry and Allied Disciplines, 45, S. 306-314.

Hacker, D. J., Bol, L., Horgan, D. D. & Rakow, E.A. (2000). Test prediction and performance in a classroom context. Journal of Educational Psychology, 92, S. 160-170.

Hadwin, A. F. & Winne, P. H. (2001). CoNoteS2: A software tool for promoting self-regulatioon and collaboration. Educational Research and Evaluation, 7, S. 313-334.

Haertel, E. H. & Lorié, W. A. (2004). Validating standards-based test score interpretations. Measurement, 2, S. 61-103.

Hamilton, L. S. (2003). Assessment as a Policy Tool. Review of Research in Education, 27, S. 25-68.

Hamilton, L. S., Klein, S. P. & Lorie, W. (2000). Using Web-Based Testing for Large-Scale Assessment. Report for the national science foundation. RAND education: Santa Monica.

Harris, D. & Khan, H. (2003). Response time to reject a takeoff. Human Factors and Aerospace Safety, 3, S. 165-175.

Harris, W. G. (2000). Best practices in testing technology: Proposed computer-based testing guidelines. Journal of e-Commerce and Psychology, 1, S. 23-35.

Hartig, J. (2003). Sensitivität für Belohnung und Bestrafung als Basis fundamentaler Persönlichkeitsdimensionen: Ein Beitrag zur Erforschung von Grays Verstärkerempfänglichkeitstheorie. Dissertation am Fachbereich Psychologie der J. W. Goethe-Universität.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In: B. Beck & E. Klieme (Hrsg.), Sprachliche Kompetenzen. Konzepte und Messung (S. 83-99). Weinheim: Beltz.

Hartig, J. & Frey, A. (2005). Application of different explanatory item response models for model based proficiency scaling. Paper presented at the 14th international meeting of the Psychometric Society in Tilburg, July 5-8 2005.

Hartig, J., Frey, A. & Jude, N. (in Druck). Validität. In: H. Moosbrugger & A. Kelava: Test- und Fragebogenkonstruktion. Berlin: Springer.

Hartig, J., Frey, A. & Ketzel, A. (2003). Modifikation des Computerspiels Quake III Arena zur Durchführung psychologischer Experimente in einer virtuellen 3D-Umgebung. Zeitschrift für Medienpsychologie, 15, S. 149-154.

Hartig, J., Jude, N. & Wagner, W. (in Druck). Methodische Grundlagen der Messung sprachlicher Kompetenzen. In: E. Klieme, W. Eichler, R. H. Lehmann, G. Nold, K. Schröder, G. Thomé & H. Willenberg (Hrsg.). Sprachliche Kompetenzen. Leistungsverteilungen und Bedingungsfaktoren. DESI-Ergebnisse Band 2. Weinheim: Beltz Pädagogik.

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In: K. Schweizer (Hrsg.,) Leistung und Leistungsdiagnostik (S. 127-143). Berlin: Springer.

Hartig, J. & Kühnbach, O. (2006). Schätzung von Veränderung mit Plausible Values in mehrdimensionalen Rasch-Modellen. In: A. Ittel & H. Merkens (Hrsg.): Veränderungsmessung und Längsschnittstudien in der Erziehungswissenschaft. Wiesbaden: Verlag für Sozialwissenschaften.

Hartinger & Fölling-Albers, M. (Hrsg.). (2004). Lehrerkompetenzen für den Sachunterricht. Probleme und Perspektiven des Sachunterrichts Bd. 14 (S. 9-18). Bad Heilbrunn: Klinkhardt.

Hathaway, S. R. & Mc Kinley, J. C. (1951). The Minnesota multiphasic personality inventory (revised). University of Minnesota, Minneapolis.

Heinssen, R. K., Glass, C. R. & Knight, L. A. (1987). Assessing computer anxiety: Development and validation of the Computer Anxiety Rating Scale. Computers in Human Behavior, 3, S. 49-59.

Helmke, A. & Hosenfeld, I. (2004). Vergleichsarbeiten – Kompetenzmodelle – Standards. In: M. Wosnitza, A. Frey & R. S. Jäger (Hrsg.), Lernprozesse, Lernumgebungen und Lerndiagnostik. Wissenschaftliche Beiträge zum Lernen im 21. Jahrhundert (S. 56-75). Landau: Verlag Empirische Pädagogik.

Herl, H. E., O'Neil, H. F., Chung, G. K. W. K. & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. Computers in Human Behavior, 15(3-4), S. 315-333.

Hölzel, B., Hartig, J., Rabl, U. & Moosbrugger, H. (2004). Experimental tests of two different explanations for item context effects in personality questionnaires. Paper presented at the VII European Conference on Psychological Assessment, Malaga, April 1st to 4th 2004.

International Test Commission (ITC) (2000). International guidelines for test use. (Version 2000). USA: Author.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. Journal of Educational and Behavioral Statistics, 25, S. 285-306.

Jude, N., Hartig, J. & Rauch, W. (2005). Erfassung von Persönlichkeitsmerkmalen im Internet und deren Bedeutung bei computervermittelter Kommunikation. In: K.-H. Renner, A. Schütz & F. Machilek (Hrsg.), Internet und Persönlichkeit (S. 119-133). Göttingen: Hogrefe.

Jung, B., Ahad, A. & Weber, M. (2005). The Affective Virtual Patient: An E-Learning Tool for Social Interaction Training within the Medical Field. In: Proceeding TESI 2005 – Training Education & Education International Conference. Nexus Media. URL: http://isnm.de/aahad/Downloads/AVP_TESI.pdf

Kindsvater, S. & Sturm, W. (2003). Computer- vs. Papier-Bleistiftvorgabe: Äquivalenzstudie zum nonverbalen Lerntest (NVLT). Zeitschrift für Neuropsychologie, 14, S. 13-21.

Klauer, K. (1986). Kriteriumsorientertes Testen: Der Schluss auf den Itempool. Zeitschrift für Differentielle und Diagnostische Psychologie, 6, S. 141-147.

Klauer, K. J. (1987). Kriteriumsorientierte Tests. Göttingen: Hogrefe.

Klieme, E. (2004a). Was sind Kompetenzen und wie lassen sie sich messen? Pädagogik, 56, S. 10-13.

Klieme, E. (2004b). Begründung, Implementation und Wirkungen von Bildungsstandards: Aktuelle Diskussionslinien und empirische Befunde. Zeitschrift für Pädagogik, 50 (5), S. 625-634

Klieme, E. (2004c). Assessment of cross-curricular problem-solving competencies. In: J. H. Moskowitz & M. Stephens (eds.). Comparing Learning Outcomes. International assessments and education policy (S. 81-107). Routledge Falmer: London and New York.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2003). Expertise zur Entwicklung nationaler Bildungsstandards. Berlin: Bundesministerium für Bildung und Forschung.

Klieme, E., Funke, J., Leutner, D., Reimann, P. & Wirth, J. (2001). Problemlösen als fächerübergreifende Kompetenz? Konzeption und erste Resultate aus einer Schulleistungsstudie. Zeitschrift für Pädagogik, 47, S. 179-200.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms bei der DFG. Zeitschrift für Pädagogik, 52, S. 876-903.

Klieme, E., Leutner, D. & Wirth, J. (Eds.). (2005). Problemlösekompetenz von Schülerinnen und Schülern. Diagnostische Ansätze, theoretische Grundlagen und empirische Befunde der deutschen PISA-2000-Studie. Wiesbaden: VS Verlag für Sozialwissenschaften.

Kobrin, J. L. & Young, J. W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration. Applied Measurement in Education, 16, S. 115-140.

Köller, Olaf (2005). Bildungsstandards – quo vadunt? Die Überprüfung und Implementierung der Bildungsstandards in Schulen. Schul-Management, 36 (6), S. 22-24.

Konak, Ö., Duindam, T. & Kamphuis, F. (2005). CITO-Sprachtest – Wissenschaftlicher Bericht. Arnhem: CITO.

Krauss, S., Kunter, M., Brunner, M., Baumert, J., Blum, W., Neubrand, M., Jordan, A. & Löwen, K. (2004). COACTIV: Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz. In: J. Doll & M. Prenzel (Hrsg.), Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung (S. 31-53). Münster: Waxmann.

Kröner, S. (2001). Intelligenzdiagnostik per Computersimulation. Münster: Waxmann.

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77, S. 1121-1134.

Laguna, K. & Babcock, R. (1997). Computer anxiety in young and older adults: Implications for human computer interactions in older populations. Computers in Human Behavior 13 (3), S. 317-326.

Lee, J. A. (1986). The effects of past computer experience on computerized aptitude test performance. Educational and Psychological Measurement, 46, S. 727-733.

Leutner, D., Klieme, E., Meyer, K. & Wirth, J. (2004). Problemlösen. In: PISA-Konsortium Deutschland (Hrsg.), PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. Münster: Waxmann.

Lienert, G. A. (1969). Testaufbau und Testanalyse (3. Auflage). Weinheim: Beltz.

van der Linden, W. J. & Glas, A. W. (Eds.) (2000). Computerized Adaptive Testing: Theory and Practice. Dordrecht; Boston; London: Kluwer Academic Publishers.

van der Linden, W. J. & Hambleton, R. K. (1997). (Eds.). Handbook of modern item response theory. New York: Springer.

Lord, F. M. (1980). Applications of item response theory to practical problems. Hillsdale, NJ: Lawrence Erlbaum.

Luecht, R. M. & Clauser, B. E. (2002). Test models for complex CBT. In: C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), Computer-based testing. Building the foundation for future assessments (S. 67-88). Mahwah, NJ: Lawrence Erlbaum.

Lumsden, J. A., Sampson, J. P. Jr., Reardon, R. C., Lenz, J. G. & Peterson, G. W. (2004). A Comparison Study of the Paper-and-Pencil, Personal Computer, and Internet Versions of Holland's Self-Directed Search. Measurement and Evaluation in Counseling and Development, 37, S. 85-94.

Maag Merki, K. (2004). Lernkompetenzen als Bildungsstandards – eine Diskussion der Umsetzungsmöglichkeiten. Zeitschrift für Erziehungswissenschaft, 7 (4), S. 539-552.

Maag Merki, K. & Grob, U. (2005). Überfachliche Kompetenzen: zur Validierung eines Indikatorensystems. In: Frey, A., Jäger, R. S. & Renold, U. (Hrsg.), Kompetenzdiagnostik – Theorien und Methoden zur Erfassung und Bewertung von beruflichen Kompetenzen (Berufspädagogik, Band 5). (S. 7-30). Landau: Verlag Empirische Pädagogik

Marsh, H. W, Trautwein, U., Lüdtke, O., Köller, O. & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: reciprocal effects models of causal ordering. Child Development, 76, S. 397-416.

Marsh, H. W, Trautwein, U., Lüdtke, O., Köller, O. & Baumert, J. (2006). Integration of multidimensional self-concept and core personality constructs: construct validation and relations to well-being and achievement. Journal of Personality, 74, S. 403-456.

Martin, R., Latour, T., Burton, R., Busana, G. & Vandenabeele, L. (2005). Covering different levels of evaluation needs by an internet-based computer-assisted testing framework for col-laborative distributed test development and delivery. Full paper submitted for presentation at ED-MEDIA 2005 – World Conference on Educational Multimedia, Hypermedia & Telecom-munications (http://www.tao.lu/downloads/publications/ Edmedia2005b.pdf).

McClelland, D. C. (1973). Testing for competence rather than for „intelligence". American Psychologist, 28, S. 1-14.

McCrae, R. R. & Costa, P. T. (1985). Updating Norman's adequate taxonomy: Intelligence and personality dimensions in natural language and in questionnaires. Journal of Personality and Social Psychology, 49, S. 710-721.

McDonald, R. P. (1997). Normal-Ogive Multidimensional Model. In: W. J. van der Linden & R. K. Hambleton (Eds.): Handbook of modern item response theory (S. 257-269). New York, Berlin: Springer

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin, 114, S. 449-458.

Messick, S. (1989). Validity. In: R. L. Linn (Ed.), Educational measurement (3rd ed., S. 13-103). New York: Macmillan.

Mikelskis, H. F. (1997). Der Computer – ein multimediales Werkzeug zum Lernen von Physik. Physik in der Schule, 35, S. 394-398.

Moosbrugger, H. & Hartig, J. (2002). Factor analysis in personality research: Some artefacts and their consequences for psychological assessment. Psychologische Beiträge, 44, S. 136-158.

Nachtigall, Kröhne, Enders & Steyer (in Druck) Causal Effects and Fair Comparison: Considering the Influence of Context Variables on Student Competencies. In: J. Hartig, E. Klieme & D. Leutner (Eds.), Assessment of Competencies in Educational Contexts. Göttingen: Hogrefe & Huber Publishers.

Nerdel, C. (2003). Die Wirkung von Animation und Simulation auf das Verständnis von stoffwechselphysiologischen Prozessen. Unveröffentlichte Dissertation: Universität Kiel.

Neuman, G. & Baydoun, R. (1998). Computerization of pencil and paper tests: When are they equivalent? Applied Psychological Measurement, 22, S. 71-83.

Normann, M., Debus, G., Dörre, P. & Leutner, D. (2004). Training of tram drivers in workload management – workload assessment in real life and in a driving/traffic simulator. In: T. Rothengatter & R. D. Huguenin (Eds.), Traffic and transport psychology – theory and application (Proceedings of the ICTTP 2000, S. 113-121). Amsterdam: Elsevier.

OECD (2001). Lernen für das Leben. Erste Ergebnisse der internationalen Schulleistungsstudie PISA 2000. Paris: OECD.

OECD (2003). The PISA 2003 assessment framework – mathematics, reading, science and problem solving knowledge and skills. Paris: OECD.

OECD (2004a). Learning for Tomorrow's World – First Results from PISA 2003. Paris: OECD.

OECD. (2004b). Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Skills from PISA 2003. Paris: OECD.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A. & Donovan, M. (1998). Conflict resolution skills assessment: A model-based, multi-media approach. Personel Psychology, 51, S. 1-24.

Page, E. B. & Petersen, N. S. (1995). The computer moves into essay grading. Phi Delta Kappan, 76 (7), S. 561-565.

Peabody, J. W., Luck, J., Glassman, P., Jain, S., Hansen, J., Spell, M. & Lee, M. (2004). Measuring the Quality of Physician Practice by Using Clinical Vignettes: A Prospective Validation Study. Annals of Internal Medicine, 141, S. 771-780.

Pinsoneault, T. B. (1996). Equivalency of computer-assisted and paper-and-pencil administered versions of the Minnesota Multiphasic Personality Inventory-2. Computers in Human Behavior, 12 (2) , S. 291-300.

Plichart, P., Jadoul, R., Vandenabeele, L. & Latour, T. (2004, November). TAO, a collaborative distributed computer-based assessment framework built on semantic web standards. Paper presented at the International Conference on Advances in Intelligent Systems – Theory and Applications AISTA, Luxembourg.

Pomplun, M., Frey, S. & Becker, D. F. (2002). The score equivalence of paper-and-and pencil and computerized versions of a speeded test of reading comprehension. Educational and Psychological Measurement, 62, S. 337-354.

Powers, D. & O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. Educational Assessment, 1 (2), S. 153-173.

Prenzel, M. & Allolio-Näcke, L. (Hrsg.). (2006). Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms. Münster: Waxmann.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J. & Schiefele, U. (Hrsg.) (2004). PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. Münster: Waxmann.

Prenzel, M., von Davier, M., Bleschke, M. G., Senkbeil, M. & Urhahne, D. (2000). Didaktisch optimierter Einsatz Neuer Medien: Entwicklung von computergestützten Unterrichtskonzepten für die naturwissenschaftlichen Fächer. In: D. Leutner & R. Brünken (Hrsg.), Neue Medien in Unterricht, Aus- und Weiterbildung. Aktuelle Ergebnisse empirischer pädagogischer Forschung (S. 113-121). Münster: Waxmann.

Rammstedt, B., Holzinger, B. & Rammsayer, T. (2004). Zur Äquivalenz der Papier-Bleistift- und einer computergestützten Version des NEO-Fünf-Faktoren-Inventars (NEO-FFI). Diagnostica, 50, S. 88-97.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Rauch, W., Hartig, J. & Moosbrugger, H. (2002). Untersuchungen zur Äquivalenz der internetbasierten und Paper-Pencil-Vorgabe eines deutschen Big-Five-Fragebogens aus dem International Personality Item Pool. Vortrag auf dem 43. Kongress der Deutschen Gesellschaft für Psychologie in Berlin vom 22. bis 26. September 2002.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In: W. J. van der Linden & R. K. Hambleton (Eds.): Handbook of modern item response theory (S. 271-286). New York, Berlin: Springer

Reips, U.-D. (2002). Standards for Internet-based experimenting. Experimental Psychology, 49 (4), S. 243-256.

Richter, T., Naumann, J. & Groeben, N. (2001). 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, 48, S. 1-13.

Ridder, H.-G., Bruns, H.-J. & Brünn, S. (2004). Online- und Multimediainstrumente zur Kompetenzerfassung. QUEM-report, Heft 86, Berlin: ESM Satz und Grafik GmbH.

Ridgeway, J. & McCusker, S. (2004). Literature Review of E-assessment. Bristol, NESTA.

Rosenblum, S., Parush, S. & Weiss, P. L. (2003a). Computerized temporal handwriting characteristics of proficient and non-proficient handwriters. American Journal of Occupational Therapy, 57, S. 139-138.

Rosenblum, S., Parush, S. & Weiss, P. L. (2003b). The in air phenomenon: Temporal and spatial correlates of the handwriting process. Perceptual and Motor Skills, 96, S. 933-954.

Rost, J. (2004). Lehrbuch der Testtheorie – Testkonstruktion (2. überarb. und erw. Aufl.). Bern: Huber.

Russell, M. K., Goldberg, A. L. & O'Connor, K. (2003). Computer-based testing and validity: A Look back into the future. Assessment in Education: Principles, Policy, and Practice, 10 (3), S. 279-93.

Rychen, D. S. & Salganik, L. H. (Eds.). (2001). Defining and selecting key competencies. Seattle: Hogrefe & Huber Publishers.

Rychen, D. S. & Salganik, L. H. (Eds.). (2003). Key competencies for a succesful life and a well-functioning society. Washington: Hogrefe & Huber Publishers.

Schermelleh-Engel, K. & Werner, C. S.. (in Druck). Methoden der Reliabilitätsbestimmung. In: H. Moosbrugger & A. Kelava: Test- und Fragebogenkonstruktion. Berlin: Springer.

Schulenberg, S. E. & Yutrzenka, B. A. (2001). Equivalence of computerized and conventional versions of the Beck Depression Inventory-II (BDI-II). Current Psychology: Developmental, Learning, Personality, Social, 20, S. 216-230.

Schuler, H. & Stehle, W. (1992). Assessment Center als Methode der Personalentwicklung. Göttingen: Hogrefe.

Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) (2007). HarmoS (http://www.edk.ch/d/EDK/Geschaefte/framesets/mainHarmoS_d.html).

Seeber, S. (2005). Zur Erfassung und Vermittlung berufsbezogener Kompetenzen im teilqualifizierenden Bildungsgang „Wirtschaft und Verwaltung" an Hamburger Berufsfachschulen. bwp@ – Berufs- und Wirtschaftspädagogik online, 8 (http://www.bwpat.de/ausgabe8/seeber_bwpat8.shtml).

Segers, M.; Dochy, F.; Cascallar, E. (Eds.) (2003): Opimising new modes of assessment: In search of q ualities and standards. Dordrecht: Kluwer.

Simonton, K. (2003). Expertise, competence, and creative ability: The perplexing complexities. In: R. J. Sternberg & E. L. Grigorenko (Eds.), The psychology of abilities, competencies, and expertise (S. 213-239). Cambridge: Cambridge University Press.

Singleton, C. (2001). Computer-based assessment in education. Educational and Child Psychology, 18 (3), S. 58-74.

Smith, B. L. (2003). Conventional versus computer-based administration of measures of cognitive ability: an analysis of psychometric, behavioural, experiential and relativity of equivalence. Unveröffentlichte Doktorarbeit. University of Wollongong (http://www-library.uow.edu.au/adt-NWU/public/adt-NWU20041006.142003/index.html).

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer/innen und das Konstrukt der diagnostischen Kompetenz. Zeitschrift für Pädagogische Psychologie, 19, S. 85-95.

Steinberg, L.S., Mislevy, R. J., Almond, R. G., Baird, A., Cahallan, C., Chernick, H., Dibello, L., Kindfield, A., Senturk, D., Yan, Duanli (2000). Using evidence-centered design methodology to design a standardsbased learning assessment. Research report, Educational Testing Service, Princeton, NJ.

Sternberg, R. J. & Grigorenko, E. (Eds.). (2003). The psychology of abilities, competencies, and expertise. New York: Cambridge University Press.

Strauß, B. & Kleinmann, M. (Hrsg.). (1995). Computersimulierte Szenarien in der Personalarbeit. Göttingen: Verlag für Angewandte Psychologie.

Streufert, S., Pogash, R. & Piasecki, M. (1988). Simulation-based assessment of managerial competence: Reliability and validity. Personel Psychology, 41, S. 537-557.

Thurstone, T. G. (1941). Primary mental abilities of children. Educational and Psychological Measurement, 1, S. 105-116.

Tousignant, M. & DesMarchais, J. E. (2002). Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: A correlation study. Advances in Health Sciences Education, 7, S. 19-27.

Troche, S., Rammstedt, B. & Rammsayer, T. (2002). Vergleich einer Papier-Bleistift- und einer computergestützten Version des Leistungsprüfsystems (LPS). Diagnostica, 48, S. 115-120.

Tseng H.-M., Tiplady, B., Macleod, H. A. & Wright, P. (1998). Computer anxiety: a comparison of pen-based personal digital assistants, conventional computer, and paper assessment of mood and performance. British Journal of Psychology, 89, S. 599-610.

Van den Branden, K., Depauw, V. & Gysen, S. (2002). A computerized task-based test of second language Dutch for vocational training purposes. Language testing, 19 (4), S. 438-452.

Vispoel, W. P. (1999). Creating computerized adaptive tests of music aptitude: Problems, solutions and future directions. In: F. Drasgow & J. B. Olson-Buchanan (Eds.), Innovations in computerized assessment (S. 151-176). Mahwah, NJ: Lawrence Erlbaum.

Wagener, D. (2001). Psychologische Diagnostik mit komplexen Szenarios. Taxonomie, Entwicklung, Evaluation. Lengerich: Pabst Science Publishers.

Walker, C. M. & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. Journal of Educational Measurement, 40, S. 255-275.

Weinert, F. E. (1999). Konzepte der Kompetenz. Paris: OECD.

Weinert, F. E. (2001a). Concept of competence: a conceptual clarification. In: D. S. Rychen & L. H. Salganik (Eds.), Defining and Selecting Key Competencies (S. 45-65). Seattle: Hogrefe & Huber.

Weinert, F. E. (Hrsg.). (2001b). Leistungsmessung in Schulen. Weinheim: Beltz.

White, R. W. (1959). Motivation reconsidered: The concept of competence. Psychological Review, 66, S. 297-333.

Wilson, M. (2005). Constructing measures. An item response modelling approach. Mawah, NJ: Lawrence Erlbaum Associates.

Winne, P. H. (1982). Minimizing the black box problem to enhance the validity of theories about instructional effects. Instructional Science, 11, S. 13-28.

Winne, P. H., Jamieson-Noel, D. L. & Muis, K. (2002). Methodological issues and advances in researching tactics, strategies, and self-regulated learning. In: P. R. Pintrich & M. L. Maehr (Eds.), New directions in measures and methods (S. 121-155). Greenwich, CT: JAI Press.

Wirth, J. (2004). Selbstregulation von Lernprozessen. Münster: Waxmann.

Wirth, J. & Klieme, E. (2003). Computer-based assessment of problem solving competence. Assessment in Education. Principles, Policy, & Practice, 10, S. 329-345.

Wise, S. L., Barnes, L. B., Harvey, A. L. & Plake, B. S. (1989). Effects of computer anxiety and computer experience on the computerbased achievement test performance of college students. Applied Measurement in Education, 2 (3), S. 235-241.

Wolf, A. (2006). Shorter Tests Through the Adaptive Use of Planned Missing Data in Sampling Designs. Unpublished PhD Thesis, Friedrich-Schiller University Jena.

Bundesministerium
für Bildung
und Forschung

20

Infolge der zunehmenden Wissensintensität in vielen Arbeits- und Lebensbereichen und der Globalisierung von Arbeits- und Bildungsmärkten wird die Frage nach der Produktivität des Bildungswesens zu einer gesellschaftlichen Kernfrage. Von der empirischen Bildungsforschung wird erwartet, dass sie diese Produktivität messbar macht. Auf Basis von Kompetenzmessungen sollen Erklärungsmodelle für Verlauf und Effektivität von Bildungsprozessen entwickelt werden sowie Interventionsstrategien wissenschaftlich untersucht werden. Der empirischen Erfassung von Kompetenzen, die sowohl als Produkt erfolgreicher Lernprozesse als auch als Voraussetzungen für erfolgreiches Lernen betrachtet werden, kommt daher eine Schlüsselfunktion für die Erforschung von Bildungsprozessen zu.

Die Nutzung technologiebasierter Erhebungsverfahren gewinnt bei der empirischen Erfassung von Kompetenzen aus mehreren Gründen an Bedeutung. Computer- und netzwerkbasiertes Assessment ermöglicht eine schnellere und ökonomischere Datenerhebung, als dies mit traditionellen Testverfahren möglich ist. Auch die Datenauswertung und die Rückmeldung von Testergebnissen werden durch technologiebasiertes Assessment erleichtert und beschleunigt. Darüber hinaus eröffnet technologiebasiertes Assessment durch die Möglichkeiten dynamischer und multimedialer Testinhalte den Zugang zu Kompetenzbereichen, die mit traditionellen Tests entweder gar nicht oder nur sehr schwer zugänglich sind.

Der vorliegende Band richtet sich an potenzielle Anwender empirischer Kompetenzerfassung in Wissenschaft und Praxis und stellt die Möglichkeiten und Chancen neuer Assessment-Technologien in diesem Bereich dar. Zugleich werden die technischen Anforderungen, die sich in spezifischen Anwendungskontexten ergeben, behandelt. Die einzelnen Beiträge des Bandes befassen sich zunächst mit konzeptionellen Grundlagen der empirischen Erfassung von Kompetenzen wie der Definition des Kompetenzbegriffs, den Qualitätsanforderungen an Testverfahren und Fragen der psychometrischen Modellierung von Kompetenzen. Hierauf aufbauend werden Möglichkeiten und potenziell kritische Aspekte technologiebasierten Assessments behandelt. Aus der Perspektive verschiedener Anwendungen werden technische Anforderungen an computer- und netzwerkbasierte Kompetenzerfassung formuliert. Schließlich werden mehrere konkrete Anwendungsszenarien technologiebasierten Assessments in verschiedenen Kontexten von Bildungsevaluation, Bildungsforschung und Lehre skizziert und die Anforderungen an die Architektur einer Software-Plattform für technologiebasierte Kompetenzdiagnostik abgeleitet.