

Testitems – Entstehung, Zweck und Rückmeldung

1 Einführung

Die Entwicklung von Testitems im Rahmen der Standards-Entwicklung in Österreich ist eines der zentralen Aufgabengebiete in der Einführung von Standards-Testungen. Um die Anforderungen, die das Kompetenzmodell an diese Aufgaben stellt, entsprechend umzusetzen wurde eine eigene Testitems-Entwicklergruppe vom BM:UKK eingesetzt. Engagierte Praktiker(innen) erstellen unter der Leitung der Test- und Beratungsstelle des Arbeitsbereiches Psychologische Diagnostik der Universität Wien, unter der Leitung von Klaus D. Kubinger, und fachdidaktischer Begleitung in einem aufwändigen, mehrstufigen Verfahren Testitems. Im folgenden Abschnitt wird versucht die Entstehung und den Zweck solcher Testitems darzulegen, sowie die Rückmeldung, die für Standards-Testungen vorgesehen ist, zu beschreiben.

2 Grundlagen

Als Grundlage für die Verankerung der Standards im Unterrichtsfach Mathematik dient das Klagenfurter Standards-Modell in der Version 4/07, welches Interessierte unter http://www.bifie.at/sites/default/files/publikationen/2007-05-09_BIST-M8.pdf (letzter Zugriff: 20.08.2008) herunterladen können. Dieses Standards-Modell ist auch für die Entwickler(innen) der Testitems sehr bedeutsam, weil verbindlich. Mit Hilfe der Testitems wird das Kompetenzmodell umgesetzt bzw. bilden die erstellten Testitems das zugrunde gelegte Kompetenzmodell genügend genau ab. Im Rahmen der Auswertung werden die Items mit Hilfe eines statistischen Verfahrens (Rasch-Modell) ausgewertet, so dass ein Stärken/Schwächen-Profil der getesteten Person erstellt werden kann. Um zu überprüfen ob ein(e)

Schüler(in) sich Wissen „nur“ für eine Prüfung angelernt hat bzw. um zu erkennen, dass Schüler(innen) den im Unterricht vermittelten Stoff tatsächlich verstanden und welche Kompetenzen sie dabei erlernt haben, müssen unterschiedliche Testitems konstruiert werden. Generell kann man jedoch sagen, dass Standards den Kernstoffbereich des Unterrichtsfaches Mathematik abdecken und mathematische Kompetenzen gemessen werden.

2.1 Tauglichkeit

Damit die von der Testitems-Entwicklergruppe erstellten Aufgaben auch tauglich – im testtheoretischen Sinn – für eine Standards-Testung sind, sollten sie, wie von Kubinger, Khorramdel et al 2007 (S. 591) beschrieben, gewissen Prinzipien der psychologischen Testtheorie sowie psychologischer Gestaltungsregeln genügen (vgl. Kubinger, Khorramdel et al 2007):

- „Items werden nur hinsichtlich gelöst oder nicht gelöst bewertet, d. h., es gibt keine irgendwie gewichteten Gutpunkte für in bestimmter Weise teilrichtige Antworten. [...]
- Items dürfen nicht aufeinander aufbauen, d. h., die Lösungen jedes einzelnen Items darf nicht davon abhängig sein, ob irgendein vorausgehendes Item gelöst oder nicht gelöst worden ist. Insbesondere soll ein und demselben Schüler zu einem Thema (zu einer inhaltlichen Angabe) nur ein einziges Item vorgegeben werden.
- Items dürfen lediglich die zu messen gesuchte Fähigkeit erfassen, d. h. zum Beispiel Spezialwissen, Arbeitsschnelligkeit, Sprachfertigkeit (bei Mathematikstandards) u. a., dürfen keinen systematischen Einfluss auf die Qualität der Testleistung haben.
- Die Lösung muss grundsätzlich eindeutig sein, d. h., alle als Lösung zu wertenden Antworten müssen ausformuliert werden und taxativ bekannt sein. Letzteres ist vor allem für Items mit so genanntem „freiem Antwortformat“ relevant.
- Die sprachliche Formulierung der Items muss extrem einfach sein; es darf nicht von der Sprachkompetenz (bei Mathematik) bzw. vom

(Fremd-)Wortschatz (bei Deutsch) u. Ä. abhängen, wie schwer die Lösung eines Items für Schüler fällt.“

2.2 Das Antwortformat

Das vorhin schon erwähnte Antwortformat ist auch bei der Entstehung von Testitems eine nicht zu unterschätzende Komponente. Bei den Testitems aus Mathematik sind verschiedene Varianten realisiert worden (vgl. Kubinger, Khorramdel et al 2007, S. 591ff):

- Freies Antwortformat
 - **mit freiem Text:**
„Die Antwort eines freien Textes wird im Zuge der Auswertung von eingeschulten Personen gemäß einem von Lehrern ausgearbeiteten Lösungsschlüssel auf Richtigkeit bewertet – aus Gründen der Ökonomie kommen solche Items relativ selten vor, jedoch bei dem Schüler mindestens einmal.“
 - **im Kästchenformat**
„Die numerische Lösung muss von dem Schüler in entsprechende (Stellenwert-)Kästchen eingetragen werden; die Anzahl der Kästchen entspricht immer genau der benötigten Anzahl.“
- Multiple-Choice Antwortformat
 - **Variante „1 aus 6“**
„Es sind sechs Antwortmöglichkeiten vorgegeben, darunter die Lösung sowie fünf Distraktoren (Antwortvorschläge, die der Lösung nahe kommen, aber eben in gewisser Weise falsch sind).“
 - **Variante „2 aus 5“**
„Es sind fünf Antwortmöglichkeiten vorgegeben, darunter befinden sich genau zwei richtige Antworten, sowie drei Distraktoren. Die Schüler werden jedes Mal darüber informiert, dass genau zwei Antworten richtig sind und beide angekreuzt werden müssen (keine zu wenig, keine zu viel), damit das Item als richtig verrechnet wird.“

Solche Testitems, die die angeführten Kriterien erfüllen, sowie mit einem entsprechenden Antwortformat versehen sind, werden zu sog. Test-Sets zusammengefasst und an das Kompetenzzentrum für Forschung und Entwicklung der PH Oberösterreich übergeben. Von dort aus erfolgt die gesamte Testadministration, wie auch die Datenverwaltung die durch die Testungen notwendig wird.

Das Antwortformat nimmt bei der Gestaltung der Testbögen eine nicht ganz unwichtige Rolle ein, da das Verhältnis im Testbogen testpsychologisch wohl überlegt sein sollte, um aussagekräftige Ergebnisse zu erhalten. Die größte Schwierigkeit in den Testbögen aus M8 ist dabei bei den ‚freien Antwortformaten‘ zu finden. Sie bereiten bei der Auswertung der Tests große Schwierigkeiten, da sich eine Normierung als schwierig erweist. Um zu sehen, welche Antworten Schüler(innen) auf Fragestellungen mit freiem Antwortformat geben, wurden Testbögen aus der Pilotphase von vier Lehrer(inne)n am Kompetenzzentrum für Forschung und Entwicklung untersucht. Die Variation dieser Antworten war entsprechend groß, was m. E. nicht weiter verwunderlich ist. Auf Anfrage am Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (bifie) wurde mir mitgeteilt, dass die prozentuelle Verteilung von offeneren Fragestellungen in einem Testbogen nicht mehr als 20% betragen darf, da insbesondere die Kosten für eine Auswertung zu hoch wären. Das Verhältnis der übrigen Antwortformate ist dem Setting für Testpsychologie überlassen. Wichtig dabei ist, dass das Verhältnis aller Antwortformate in einem testpsychologisch aussagekräftigen Verhältnis steht. Nach Auskunft von Frau Khorramdel von der Test- und Beratungsstelle sieht das Testsetting, also die Verteilung der Antwortformate, wie folgt aus:

1. Schwierigkeitsgrad des Beispiels: Ausschlaggebend ist der Itemleichtigkeitsparameter bei erfolgter Rasch-Modell-Analyse ansonsten die Einschätzung der Itementwickler(innen)gruppe. Für jede Leistungsgruppe in der Hauptschule (I, II, III) müssen nämlich unterschiedlich schwere Testhefte erstellt werden.

2. Inhaltliche Festlegung: bei den verwendeten Beispielen wird darauf geachtet, dass ein bestimmtes Thema (z. B. Fußball) nicht mehrmals in einem Testheft vorkommt
3. Art der Antwortformate: Antworten im Kästchenformat bzw. im Multiple-Choice-Format werden auf alle Testhefte gleich verteilt, d. h. in jedem Testheft finden sich gleich viele Antwortformate für Kästchen bzw. Multiple-Choice, der Rest (ca. 1 – 2 Aufgaben) sind freie Antwortformate.

Aufgrund der Tatsache, dass nicht allzu viele Aufgaben mit freiem Antwortformat vorhanden sind, ist das beschriebene testpsychologische Testsetting i. Allg. für Mathematikdidaktiker(innen) nicht weiter „anstößig“. Es sollte jedoch darauf geachtet werden, dass mehr freie Antwortformate, insbesondere das Format mit „freiem Text“ in die (zukünftigen) Tests Einzug halten, da mathematisches Verständnis umfassender und genauer überprüft werden kann, auch wenn die Korrektur solcher Aufgaben aufwändiger ist.

2.3 Testung und fachdidaktische Relevanz

Als die Entwicklung der Bildungsstandards für die Sekundarstufe I aufgenommen wurde, hat eine rechtliche Verankerung seitens des Gesetzgebers nicht stattgefunden. Dies gesetzliche Lücke wurde jedoch im Kalenderjahr 2008, mit Juli, behoben (§17 SCHUG) und die Bildungsstandards, auch aus Mathematik, in den der Grundschule und Sekundarstufe I (Hauptschule und AHS-Unterstufe) gesetzlich verbindlich eingeführt; per Erlass werden die Schulen davon in Kenntnis gesetzt; die Testung wird durch Verordnung geregelt.

Um eine Möglichkeit zur Validierung der Testitems zu erhalten, wurden bis einschließlich dem Kalenderjahr 2008 Feldtests mittels Pilot-Testungen durchgeführt. Mit Hilfe dieser Pilot-Testungen erhoffte man sich Antworten auf Fragen, wie z. B. (1) inwieweit kann der Entwicklungsstand der Mathematik-Standards und die Bedingungen ihrer Implementierung die Voraussetzungen für eine zunehmend positive Unterrichtskultur erfüllen, (2) auf welche

Problem- und Entwicklungsbereiche ist besonders zu achten, (3) wo bedarf es der weiteren Klärung von Rahmenbedingungen für die Praxis?

Bei den Fragen der Implementierungen sowie Klärung der Rahmenbedingungen für die Praxis ist in jedem Fall auch die Mathematik Fachdidaktik gefragt. Kritisch ist im beschriebenen Modell der Durchführung daher anzumerken, dass die fachdidaktische Begleitung – wie bereits im Einleitungskapitel 1 angeführt – neben den Testpsycholog(inn)en ausschließlich von engagierten Praktiker(inne)n eingebracht wird. Erfahrungsgemäß verfügen insbesondere Praktiker(innen) zwar über ein großes Methodenrepertoire, die Kenntnisse über fachdidaktische Modelle und Prinzipien sind aber zumeist nur sehr dürftig bzw. überhaupt nicht vorhanden. Über Erfahrungen dazu berichtet Edith Schneider: „...Zahlreiche Erfahrungen und Untersuchungen zeigen, dass ... das vermittelte theoretische Wissen von den Praktiker(inne)n kaum mit deren Praxis in Beziehung gesetzt wird bzw. die vermittelten Theorien kaum oder nicht in der intendierten Form in deren Praxis zum Einsatz kommen ...“ (Schneider 2004, S. 1). Einem ausgewogenen Verhältnis von Theorie und Praxis in der Fachdidaktik wurde hier nicht wirklich Rechnung getragen.

Die Pilottestungen fanden in ausgewählten „Pilotschulen“ statt, die sich freiwillig dazu bereit erklärt haben, das Projekt Bildungsstandards zu unterstützen. Die Pilottestung, insbesondere im Unterrichtsfach Mathematik, diente vor allem zur Überprüfung und Verbesserung der entwickelten Testverfahren und Testitems.

Trotz dieses Versuchscharakters der Testungen erhielten die getesteten Schüler(innen), die davon betroffenen (Klassen-)Lehrer(innen), sowie die Schulaufsicht (Schulleiter(innen), Schulbehörde) eine ausführliche Rückmeldung über die Testergebnisse. Der/die Schüler(in) erhält eine individuelle Rückmeldung über seine/ihre Stärken und Schwächen, alle anderen Beteiligten Personen eine kumulierte Rückmeldung.

Für das Jahr 2009 sind „Baseline-Testungen“ vorgesehen. Bei diesen wird eine repräsentative Auswahl an Schüler(inne)n der Sekundarstufe I im Unterrichtsfach Mathematik (sowie den anderen betroffenen Fächern) getestet. Die dabei gewonnenen Daten sollen als Vergleichsbasis für zukünftige Testungen dienen, da die Daten der Pilottestungen lediglich Informationscharakter hatten.

3 Rasche Verträglichkeit oder Rasch-Verträglichkeit?

Grundlage für die Konstruktion der heutzutage gebräuchlichen Testverfahren ist die Klassische Testtheorie. Diese beruht jedoch auf Axiomen, d. h. grundsätzlichen Annahmen, die nicht beweisbar sind. Daher ist ein solches Testverfahren für die Analyse von Standards-Beispielen nicht geeignet, da man mit Hilfe der klassischen Testtheorie nicht beantworten kann, ob alle Items dieselbe latente Variable, d. h. das vermutete Persönlichkeitskonstrukt, messen. Um eine Vergleichbarkeit zwischen den Testpersonen zu erzielen, setzt man daher ein Testverfahren mit probabilistischem Ansatz ein, welches meistens mit dem Konzept der spezifisch objektiven Vergleiche verbunden ist. Ein solches Testverfahren ist das Rasch-Modell (benannt nach dem Begründer des Modells Georg Rasch, der 1960 erstmals Ansätze davon publizierte (Rasch 1960, 1968), mit dessen Hilfe man in der Testpsychologie Informationen über eine Person erhält, in dem man die Anzahl der von der Person in einem Test behandelten (gelösten) Items betrachtet. Der Psychologe Fischer beschreibt dieses Verfahren (Fischer 1988, S. 88): „Der Rohwert, definiert als Anzahl der richtigen Antworten einer Person, ist eine erschöpfende Statistik für ihren Personenparameter, sofern die Itemparameter bekannt sind; die Itemparameter (und im Prinzip auch die Personenparameter) lassen sich mittels bedingter Maximum-Likelihood-Methoden schätzen; die Modellanpassung sowie weiterführende Hypothesen können mittels bedingter Inferenzmethoden geprüft werden; für die Modellanwendung genügen vergleichsweise kleinere Stichproben.

All diese günstigen Eigenschaften werden häufig als Gründe für die Verwendung des Rasch-Modells angeführt. Darüber hinaus spielt aber auch

die ‐Spezifische Objektivit t‐ der Ergebnisse eine wichtige Rolle f r seine Rechtfertigung. Grob gesprochen versteht man darunter, da  der Vergleich zweier Personen auf dem latenten Kontinuum in einem bestimmten Sinne unabh ngig von den Itemparametern ist, sofern beiden Personen dieselben Items vorgeben wurden, bzw. da  der Vergleich zweier Items in analoger Weise unabh ngig von den Personenparametern der getesteten Personenstichprobe ist.‐

In diesem Rasch-Modell ist es also nicht mehr wichtig welche Beispiele eine Person in einem Test l st, sondern wie viele Beispiele besagte Person l sen konnte. Das Rasch-Modell beschreibt daher die Wahrscheinlichkeit, dass eine Testperson ein Item in Abh ngigkeit eines Personenparameters l st. Dabei wird der Personenparameter durch die wahre F higkeit der Testperson und die wahre Schwierigkeit des Items (= Itemparameter) festgelegt. Konkret bedeutet das, dass eine F higkeit einer Person nicht determiniert, ob die Testperson bei einem Item mit einer gewissen Schwierigkeit zu einer L sung kommt, sondern nur probabilistisch in der Hinsicht beschreibt, als die Realit t der Ergebnisse f r gr o ere Testpersonen-Stichproben bei konstantem Schwierigkeitsgrad gr o er wird. Knoche & Lind 2000 (S. 8–27) haben sich aus fachdidaktischer und fachlicher Sicht ausf hrlich mit dem Rasch-Modell auseinandergesetzt, da dieses Test-Modell bereits in der TIMSS-Studie Verwendung gefunden hat. Die beiden Autoren (Knoche & Lind 2000, S. 4) f hren in das Modell wie nachfolgend zitiert ein:

„Bei der Modellierung von Testsituationen kann man unterstellen, dass die Bearbeitung einer Testaufgabe unter Zeitdruck ein Zufallsexperiment ist, in dem der Proband und die Aufgabe zusammen ein *Zufallsger t* bilden, das die *Bewertung* der gezeigten Reaktionen als Ergebnis hat. Im Gegensatz zu mechanischen Zufallsger ten kann man jedoch diesen Versuch in den meisten F llen nicht wiederholen und muss daher Annahmen  ber *quantifizierbare* Aufgabeneigenschaften und Probandeneigenschaften machen, die eine Sch tzung damit definierter Parameter aus den Testergebnissen einer Probanden*population* erlauben.‐

Das Rasch-Modell wird mathematisch ausf hrlich erl utert und auch kritisch gegen ber diesem Modell Stellung bezogen, denn „ ... zur

Prüfung didaktischer Hintergrundtheorien ist ein solches Vorgehen allerdings nicht angebracht.“ (vgl. Knoche & Lind 2000, S. 7)

Die Analyse der von der Gruppe erstellten Testitems wurde vom Team Testpsychologie der Universität Wien durchgeführt. Die Bedeutung des verwendeten (Rasch-)Modells ist für die (psychologischen) Diagnostiker darin begründet, dass man in diesem Modell die Verrechnung der Testleistungen so vorschreibt, dass als Testwert nur die Anzahl gelöster Aufgaben bestimmt wird, d. h. dieser Wert unabhängig davon ist, welche Aufgaben (nicht) gelöst wurden – daher müssen diese Aufgaben des Tests notwendigerweise dem Rasch-Modell folgen. So kann garantiert werden, dass der Testwert tatsächlich die gesamte relevante Information in Bezug auf die fragliche Fähigkeit der getesteten Person(en) ausschöpft (vgl. Fischer 1974).

In zahlreichen Analysen hat sich gezeigt, dass die Handlungsdimensionen im Kompetenzmodell M8, ebenfalls wie in M4, dem Rasch-Modell besser genügen. Nur diese Achse wird daher der Rasch-Modell-Analyse unterzogen.

Aus (fach-)didaktischer Sicht ergibt dies ein verzerrtes Bild, da die im Standards-Modell wesentliche Inhaltsdimension nicht in den Testleistungen abgebildet wird und auch die Komplexitätsdimension keinerlei Einfluss auf das Ergebnis des Standards-Tests hat.

3.1 Der Aufgabenpool

Im Zeitraum 2005 bis 2007 entwickelte die Item-Entwickler(innen)gruppe Beispiele die für Testungen verwendet wurden. Die Anzahl der entworfenen Beispiele übersteigt jedoch die Zahl der in den Testheften verwendeten Beispiele, daher konnten noch nicht alle Aufgaben getestet und einer Rasch-Modell-Analyse unterzogen werden. Im Aufgabenpool existieren zwei Einteilungen für die vorhandenen Beispiele – „Rasch-Modell-konform“ und „testreif“. Die Kategorie „testreif“ ist für ein Beispiel dann anzuwenden, wenn es

von der Gruppe für eine Testung freigegeben wurde, diese Aufgabe jedoch noch nicht im Feldtest eingesetzt wurde bzw. es keine empirischen Daten über dieses Beispiel vorliegen. Die Freigabe dieser Aufgaben konnte nur dann erfolgen, wenn sie den Anforderungen vorhandener Regelkataloge (siehe Grundlagen, Antwortformat, sprachliche Formulierung etc.) genügte. Für den Testitempool M8 ergibt sich folgendes Bild (vgl. Kubinger, Khorramdel et al 2007, S. 595):

Handlungsdimension	H1 Darstellen, Modell- bilden	H2 Operieren, Rechnen	H3 Interpretieren, Dokumen- tieren	H4 Argumen- tieren, Begründen
Rasch- Modell konform	47	45	39	21
testreif	43	49	48	32

Tabelle 1. Anzahl der testfertigen Beispiele im geschlossenen Aufgabenpool M8

Bezogen auf die Handlungsdimensionen des Kompetenzmodells ergaben sich also bis Ende 2007 in Summe 152 Aufgaben die dem Rasch-Modell genügte, sowie 172 Aufgaben, die als testreif eingestuft wurden, jedoch noch nie in einer Testung zum Einsatz kamen. Die Entwicklung dieser Aufgaben erfolgte jedoch nicht unter Berücksichtigung des Kompetenzmodells Version 4/07, welches die (gesetzliche) Grundlage für die M8-Standards darstellt. Daher ergaben sich noch massive Umschichtungen von Beispielen in den jeweiligen Handlungsdimensionen, bis hin zur Aussortierung von Beispielen aus dem Testitempool. Die Anzahl von (in Tabelle 1 angegebenen) 324 testfertigen Beispielen wird sich weiter verändern, da die Item-Entwickler(innen)gruppe weitere Beispiele 2008 in den Pool aufgenommen hat. Genaue Zahlen darüber liegen jedoch (noch) nicht vor; derzeit gibt es ca. 300 testreife Beispiele, ca. 150 davon sind mit dem Rasch-Modell analysiert. Das ursprüngliche Ziel von ca. 500 Aufgaben ist damit noch nicht erreicht.

3.2 Exemplarische Aufgabenstellung

Um sich die auf die Testitems angewandte Rasch-Modell-Analyse besser vorstellen zu können möchte ich dies exemplarisch anhand einer Aufgabenstellung welche freigegeben wurde, d. h. also **nicht** mehr im Testitempool zu finden ist, vorstellen (Quelle: www.bifie.at – letzter Zugriff: 20.08.2008):

Von einem Rechteck sind der Flächeninhalt $A = 49 \text{ cm}^2$ und der Umfang $u = 32 \text{ cm}$ gegeben.

Handelt es sich um ein Quadrat?

Schreib deine Begründung im Antwortbogen auf!

Aufgrund der sich ständig ändernden Zahlen in der Rasch-Modell-Analyse und auch aufgrund rechtlicher Grundlagen, darf kein Zahlenmaterial über die Rasch-Modell-Analyse veröffentlicht werden. Darum bilde ich hier auch nur den möglichen Raster zur Analyse eines Beispiels ab.

Knoten		Bezeichnung		Antwortformat		Schwierigkeits-einschätzung		
H4 / I3		Rechteck oder Quadrat		frei		2. LG		
be- arbeitet	nicht be- arbeitet	gelöst	nicht gelöst	Anteil: Anz. nicht be- arbeitet	Anteil: gelöst	Anteil: nicht gelöst	Item- leichtig- keits- parameter	Fähigkeits- parameter

Tabelle 2. Rasch-Modell-Analyse-Raster

Dieses Beispiel hat dem Rasch-Modell genügt. Aber wie interpretiert man nun die in Tabelle 2 dargestellten Kategorien? Der Knoten ordnet das Beispiel in das Kompetenzmodell ein; H4 steht dabei für „Argumentieren, Begründen“ und I3 für „Geometrische Figuren und Körper“; dieses Beispiel zielt also darauf ab, dass Schüler(innen) die Fähigkeit besitzen im Bereich der ebenen Figuren zu argumentieren bzw. Begründungen zu geben. Die Bezeichnung bezieht sich auf den Titel der Aufgabe; das Antwortformat – hier: frei – kann nur die in 2.2 beschriebenen Ausformungen besitzen; die Schwierigkeits-

einschätzung wird von den Itemersteller(inne)n selbst getroffen, die endgültige Festlegung erfolgt erst nach einer ausführlichen Diskussion durch die Gruppe. Für alle in der Entwicklung mit dieser Aufgabe befassten Personen ist zudem noch ersichtlich, in welchen Testheften dieses Beispiel eingesetzt wurde. Die empirischen Daten, welche durch die Testdurchführung gewonnen wurden, sind vor allem für das Team Testpsychologie sehr interessant. Für eine Aufgabe kann man erkennen, dass eine bestimmte Anzahl von Schüler(inne)n der Sekundarstufe I (AHS Unterstufe, HS 1. und 2. LG) diese Aufgabe in einem der Testhefte vorgelegt bekommen haben, eine gewisse Anzahl davon haben diese dann nicht bearbeitet. Von allen getesteten Schüler(inne)n konnte wiederum eine bestimmte Anzahl die Aufgabe richtig lösen, eine bestimmte Anzahl erarbeitete eine falsche Lösung für dieses Beispiel. Die nachfolgenden Anteile sind Prozentwerte gerundet auf zwei Dezimalstellen; d. h. ein gewisser Prozentsatz der „getesteten“ Schüler(innen) hat dieses Beispiel nicht bearbeitet, ein bestimmter Prozentsatz hat die Aufgabe gelöst und ein weiterer Prozentsatz hat die Aufgabe falsch bzw. nicht gelöst (d. h. die Aufgabe wurde begonnen, aber nicht fertiggestellt). Der Itemleichtigkeitsparameter, ein geschätzter Wert, normalerweise zwischen -4 und 4 (extremere Werte sind möglich) ergibt sich durch das Rasch-Modell, indem die vorhandenen Werte eingesetzt werden. Mit Hilfe dieses Parameters kann man die Aufgaben kategorisieren, d. h. wie leicht ist die Aufgabe aufgrund der Analyse tatsächlich; aufgrund dieser Kategorisierung kann sie später wieder entsprechend eingesetzt werden. Der Personenparameter zur Messung der Fähigkeit von Schüler(inne)n ist ebenfalls ein wichtiger Wert zur Kategorisierung eines Beispiels. Dieser Parameter wird dazu verwendet um den/die getestete(n) Schüler(in) rückschließen zu lassen, wie er/sie das Beispiel gelöst hat; aber auch um eine Vergleichbarkeit innerhalb einer Klasse zu ermöglichen, wird dieser Wert verwendet; bzw. um getestete Klassen untereinander zu vergleichen, kann dieser Fähigkeitsparameter herangezogen werden. Dieser Wert ist durch ein individuelles Konfidenzintervall angegeben, das eng wird, wenn für die jeweilige

Personenfähigkeit mehrere Items Informationen liefern bzw. das breit wird, wenn wenig Items für diesen Bereich Informationen liefern.

Interessant sind auch Kriterien, warum die Beispiele dem Rasch-Modell nicht genügen können. Dafür hat das Team Testpsychologie mehrere Einteilungen:

- Teilungskriterium Geschlecht (weiblich, männlich)
- Teilungskriterium Muttersprache (deutsch, nicht-deutsch)
- Teilungskriterium Region (Ost- Westösterreich)
- Teilungskriterium Median (leistungsstarke- bzw. schwache Schüler(innen))

Ergeben sich in einem der vier angeführten Teilungskriterien zu große Unterschiede in den Kriterien, so ist das Beispiel nicht mehr Rasch-Modell konform und muss um in weiteren Testungen Einsatz zu finden, überarbeitet werden, für testreif erklärt werden und dann nochmals die Rasch-Modell-Analyse durchlaufen.

4 Ergebnisrückmeldung

Seit dem Jahr 2005 sind in Österreich jährlich Pilottestungen für die Bildungsstandards aus Mathematik durchgeführt worden. Die Anzahl der an den Testungen teilnehmenden Schüler(innen) hat von Jahr zu Jahr zugenommen. Die Dauer eines solchen Tests beträgt 80 Minuten. Die Durchführung erfolgt durch spezielle Testadministrator(inn)en. Die gängige Praxis (zumindest in der Pilotphase) ist/war, freiwillige motivierte Lehrer(innen) dafür zu rekrutieren, die dann die Pilotschulen mit den Testbögen aufsuchten und dort die Testungen durchführten. Ausgestattet mit einem Heft „Instruktion zur Testadministration der Standard-Tests – Überprüfung der Bildungsstandards für Mathematik und Deutsch am Ende der 4. und 8. Schulstufe“ aus der Skriptenreihe der Test- und Beratungsstelle, sowie einem halbtägigen Seminar zur Einführung in die Bildungsstandards waren dies die „geschulten Lehrer aus jeweils anderen Schulen“ (vgl. Kubinger, Khorramdel et al 2007, S. 596), die die Testungen

durchführten. Danach mussten diese Testadministrator(inn)en alle mitgebrachten Unterlagen an das vorhin schon erwähnte Kompetenzzentrum für Forschung und Entwicklung der PH Oberösterreich postalisch übermitteln, wo die Testbögen (automatisiert) ausgewertet werden/wurden. Bei der automatischen Auswertung der Testbögen ergeben sich jedoch große Schwierigkeiten mit dem freien Antwortformat, da diese Auswertung nur von geschultem Personal erfolgen kann. Eine Lösung dieses Problems scheint im Moment noch nicht zu existieren. Gerade deswegen wird dieses Antwortformat in den Testbögen, wie in 2.2 beschrieben, nicht sehr oft eingesetzt werden.

Wurde eine Schule als Testschule ausgewählt, so erhalten alle an dieser von der Testung betroffenen Personen Rückmeldung über die Ergebnisse des Tests – Schüler(innen), Lehrer(innen) der getesteten Klassen, Direktor(inn)en der betroffenen Schule, sowie Landesschulinspektor(inn)en, die für diese Schule zuständig sind. Die Ergebnisse stehen online, verschlüsselt zum Abruf unter <http://www.bildung-standards.at> (letzter Zugriff: 20.08.2008) bereit.

Bemerkenswert ist die Einführung, in der darauf hingewiesen wird, dass nicht alle Beispiele ausgewertet werden, sondern lediglich die Aufgaben, die bereits testtheoretisch erprobt wurden. Die Rahmenbedingungen der Pilottestung – Testzeit von 80 Minuten, 30 Aufgabenstellungen – werden in den (regulären) Standards-Testungen genauso auftreten.

Folgt der/die Lehrer(in) dem Link so wird er zu einer neuen Maske geführt, in der er/sie alle für ihn/sie interessanten Ergebnisse abrufen kann. Der Vergleich zu andern AHS-Klassen in Österreich kann aufgrund der flächendeckend vorhandenen Pilotschulen durchgeführt werden. Im Schuljahr 2008/09 wird eine Baseline-Testung österreichischer Schüler(innen) durchgeführt werden. Die hier erhaltenen Ergebnisse werden dann als Vergleichswerte herangezogen werden.

In der Rückmeldemaske können Lehrer(innen) außerdem die Ergebnisse ihrer Schüler(innen) in den Kompetenzbereichen abrufen.

Folgen Lehrer(innen) einem der vorhandenen Links, erhalten sie eine grafische Auswertung der durchschnittlichen Fähigkeit der überprüften Klasse im Vergleich zu anderen österreichischen AHS-Klassen. Die Testleistungen können auch detailliert in den vorhandenen Kompetenzbereichen, wieder im Vergleich zu anderen AHS-Klassen in Österreich, betrachtet werden.

Eine ähnliche Rückmeldung wie Lehrer(innen), lediglich allgemeiner – weil auf die jeweilige Schule im Vergleich zu anderen Schulen bezogen, erhalten Direktor(inn)en. Dabei ist auch eine verfeinerte Darstellung in den jeweiligen Kompetenzbereichen möglich. Der Vorteil dieses Systems ist darin zu finden, dass die Direktor(inn)en keine Rückschlüsse auf einzelne Lehrer(innen) durchführen können, sondern lediglich erkennen, wo ihre Schule im Vergleich zu den anderen österreichischen Schulen liegt.

Aufgrund dieser Rückmeldung sind Direktor(inn)en dann angehalten Maßnahmen zur Qualitätssicherung in ihrer Schule zu setzen. Diese müssen zuvor aber mit dem gesamten Mathematik-Lehrer(innen)-Team besprochen und können dann erst durchgeführt werden.

5 Ausblick

Seit dem Beschluss und der Absicht, Bildungsstandards in Österreich für das Unterrichtsfach Mathematik einzuführen, ist sehr viel Entwicklungsarbeit, nicht zuletzt in Testitems, gesteckt worden. Man darf aber, egal wie man über dies neue eingeführte Bildungsmonitoring denken mag, nicht übersehen, dass in den kommenden Jahren weiterhin noch sehr viel an Entwicklungsarbeit geleistet werden muss. Viele der entwickelten Konzepte sind gerade durch den Feldtest nochmals überarbeitet und verbessert worden und gerade deswegen ist es auch notwendig, dass sich Fachdidaktiker(innen) in die Entwicklung von Bildungsstandards für das Unterrichtsfach Mathematik mit

einbringen und entscheidend mitwirken. Anhand der ersten großen Standards-Testung, der sog. Baseline-Testung, im Schuljahr 2008/09 wird sich zeigen, welche (zusätzlichen) Schwierigkeiten eventuell noch auftreten können bzw. werden. Vor allem die Auswertung der Standards-Bögen und die zeitgerechte Rückmeldung für alle Beteiligten wird eine große Herausforderung werden. Es werden sich aber sicherlich weitere wissenschaftliche Fragestellungen für Fachdidaktiker(innen) aus Mathematik auftun. So wurde bei den Auswertungen der bisher durchgeführten Pilottestungen noch nicht darauf geachtet, ob österreichische Schüler(innen) in bestimmten Kompetenzbereichen besondere Stärken oder auffällige Schwächen besitzen. Das Team Testpsychologie war vertraglich nur dazu verpflichtet, die Feldtests entsprechend vorzubereiten und die Auswertung und Evaluation durchzuführen. Untersuchungen im Detail sind bis dato noch nicht durchgeführt worden und sind für Fachdidaktiker(innen) aus Mathematik sicherlich eine interessante Herausforderung – v. a. wenn man die Ergebnisse möglicherweise mit den vorhandenen PISA-Daten vergleichen kann.

Eine weitere interessante Fragestellung, nicht nur für Fachdidaktiker(innen) aus Mathematik, könnte sich aber auch dahingehend ergeben, als man das verwendete Analyseverfahren (Rasch-Modell) hinterfragt. Es gibt eine Reihe von Testverfahren aus der Testpsychologie (vgl. Kubinger 1988) die man zur Analyse der Beispiele heranziehen kann und die u. U. ganz andere Ergebnisse bei der Analyse liefern würden. Aber auch die isolierte Betrachtung der Handlungsdimensionen ist nicht notwendigerweise ein Idealzustand. Ein Analyseinstrument zu finden oder zu entwickeln, mit dessen Hilfe man die im Kompetenzmodell vorhandenen Dimensionen verknüpfen kann und die ansprechende Ergebnisse liefert, kann eine weiterführende Forschungsfrage für die (österreichische) Wissenschaft sein. Dazu bedarf es aber auch großer Unterstützung seitens der österreichischen Politik und des bifie.

Einige Änderungen werden sich aber auch für Lehrer(innen) ergeben. So werden die Bildungsstandards aus dem Unterrichtsfach Mathematik

die österreichische Schulbuchlandschaft, aber auch die Fortbildungslandschaft nachhaltig beeinflussen. Unterrichtende Lehrer(innen) werden hier gezielt geschult werden müssen, um mit den kommenden Anforderungen auch gut zu Recht zu kommen. Dies kann eine zukünftige Aufgabe der österreichischen Fachdidaktik sein, sich hier entsprechend einzubringen und gezielte Maßnahmen nicht nur zu treffen, sondern auch zu setzen.

Änderungen werden sich aber natürlich auch für Schüler(innen) und die österreichische Bildungslandschaft ergeben, denn die Bildungsstandards aus Mathematik für die 8. Schulstufe sind das Grundgerüst für weitere Maßnahmen im österreichischen Schulwesen. So wird die Entwicklung der Bildungsstandards aus Mathematik in absehbarer Zeit in der Konzeption einer zentralen schriftlichen Reifeprüfung enden. Das entsprechende Projekt ist vom Ministerium für Unterricht, Kunst und Kultur (BM:UJK) bereits genehmigt und österreichische Fachdidaktiker(innen) arbeiten bereits intensiv daran.

Literatur

- Fischer, G. H. (1974): *Einführung in die Theorie psychologischer Tests*. Bern: Huber.
- Fischer, G. H. (1988): *Spezifische Objektivität: Eine wissenschaftstheoretische Grundlage des Rasch-Modells*. In: Kubinger, K. D. (Hrsg.): *Moderne Testtheorie – ein Abriss samt neuesten Beiträgen*. München, Weinheim: Beltz Test Gesellschaft, Psychologie Verlags Union, S. 87-111.
- Knoche, N. & Lind, D. (2000): *Eine Analyse der Aussagen und Interpretationen von TIMSS unter Betonung methodologischer Aspekte*. In: JMD 21, Heft 1, S. 3-27.
- Kubinger, K. D. (1988): *Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie*. In: Kubinger, K. D. (Hrsg.): *Moderne Testtheorie – ein Abriss samt neuesten Beiträgen*. München, Weinheim: Beltz Test Gesellschaft, Psychologie Verlags Union, S. 19-83.
- Kubinger, K. D., Khorrandel, L. et al (2007): *Large-Scale Assessment zu den Bildungsstandards in Österreich: Testkonzept, Testdurchführung und*

- Ergebnisverwertung*. In: *Erziehung & Unterricht*, 7–8/2007, 157. Jahrgang, S. 588-600.
- Rasch, G. (1960): *Probabilistic models for some intelligence and attainment test*. Copenhagen, Paedagogiske Institute.
- Rasch, G. (1968): *A mathematical theory of objectivity and its consequences for model construction*. Paper presented at the European Meeting on Statistics Econometrics and Management Science. Amsterdam, S. 2-7.
- Schneider, E. (2004): *Professionalität von Lehrerinnen und Lehrern*. In: *ZDM*, Vol. 36 (1), S. 1-2.
- Standards für die mathematischen Fähigkeiten österreichischer Schülerinnen und Schüler am Ende der 8. Schulstufe*. Version 4/07. Herausgegeben vom Institut für Didaktik der Mathematik – Österreichisches Kompetenzzentrum für Mathematikdidaktik – der Universität Klagenfurt, Klagenfurt 2007. http://www.bifie.at/sites/default/files/publikationen/2007-05-09_BIST-M8.pdf (letzter Zugriff: 08.09.2008).