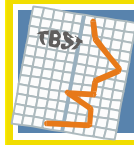




universität
wien



**Test- und
Beratungsstelle**

Fakultät für Psychologie
Arbeitsbereich Psychologische Diagnostik

Standard-Tests zu den Bildungsstandards in Österreich

**Wissenschaftlicher Hintergrund und Hinweise zur Interpretation
der Ergebnisse der Standard-Tests**

**Klaus D. Kubinger
Martina Frebort
Stefana Holocher-Ertl
Thomas Pletschko**

DAS ZUKUNFTSMINISTERIUM

bm:bwk

© *Test- und Beratungsstelle*, Arbeitsbereich Psychologische Diagnostik: www.testzentrum.at
Fakultät für Psychologie der Universität Wien, Liebiggasse 5, A-1010 Wien
gem. Mediengesetz für den Inhalt verantwortlich: Univ. Prof. Dr. Mag. Klaus D. Kubinger

© 2006, Version 1.1

Herstellung: bmbwk

Inhaltsverzeichnis

1. Bildungsstandards – Einleitung und Ziele	4
1.1. Was sind Bildungsstandards?	4
1.2. Wozu dienen Bildungsstandards?	4
1.3. Wer entwickelt Bildungsstandards? Wo werden sie eingesetzt?.....	5
1.4. Wie sind die Bildungsstandards aufgebaut?	5
1.5. Ab wann gelten die Bildungsstandards? Wie wird das Erreichen der Standards überprüft?.....	7
1.6. Was geschieht mit den Ergebnissen?.....	7
1.7. Vergleich von PISA und Bildungsstandards.....	8
2. Prinzipien bei der Erstellung der Standard-Tests.....	9
2.1. Testtheoretische Prinzipien bei der Itemkonstruktion zu den Unterrichtsfächern Mathematik und Deutsch	10
2.2. Antwortformate.....	11
2.2.1. <i>Freies Antwortformat mit freiem Text</i>	12
2.2.2. <i>Freies Antwortformat im Kästchenformat</i>	13
2.2.3. <i>Multiple-Choice „1 aus 6“</i>	13
2.2.4. <i>Multiple-Choice „2 aus 5“</i>	14
3. Ergebnismrückmeldung	14
3.1. Ergebnismrückmeldung auf Schölerebene	14
3.1.1. <i>Generell</i>	14
3.1.2. <i>Vierte Schulstufe</i>	16
3.2. Ergebnismrückmeldung auf Lehrerebene	17
3.3. Ergebnismrückmeldung auf Schulleiterebene.....	21
3.4. Ergebnismrückmeldung auf Schulaufsichtsebene	22
4. Grenzen der Interpretierbarkeit	22
5. Verwertung der Testergebnisse	23
6. Frequently Asked Questions	24
Literatur.....	25
Anhang – Rasch Modell	26
Organigramm: Struktur der kooperierenden Teams	31

Vorbemerkung: Der besseren Lesbarkeit wird im Folgenden stets von „dem Schüler“, „dem Lehrer“ und ähnlichem gesprochen; selbstverständlich sind Schülerinnen, Lehrerinnen usw. stets mit gemeint.

1. Bildungsstandards – Einleitung und Ziele^{1, 2, 3}

1.1. Was sind Bildungsstandards?

Bildungsstandards in Österreich sind als Regelstandards (nicht als Mindeststandards) konzipiert und legen fest, welche Kompetenzen Schüler bis zu einer bestimmten Schulstufe nachhaltig erworben haben sollen – unter „Kompetenzen“ kann man verstehen: „die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können“ (Weinert, 2001).

Die Bildungsstandards konzentrieren sich dabei auf die Kernbereiche eines Unterrichtsfaches und beschreiben die erwarteten Lernergebnisse, wobei fachliche Basisqualifikationen definiert werden, die für die weitere schulische Bildung bzw. berufliche Ausbildung von Bedeutung sind. Bildungsstandards konkretisieren eine normative Erwartung, auf die die Schule hinarbeiten soll (sie legen allerdings nicht fest, was guter Unterricht ist und sie sind auch kein Instrument für ein Qualitätsranking²).

Besonders wichtig ist die Feststellung, dass Bildungsstandards nicht das Lehren und Lernen und damit auch nicht den Prozess der schulischen Bildung standardisieren. Z.B. die Erweiterungsgebiete in den Lehrplänen der Sekundarstufe I bleiben von den Bildungsstandards ausgenommen.

1.2. Wozu dienen Bildungsstandards?

Bildungsstandards sollen Lehrern bessere Orientierung und mehr Sicherheit in ihrer unterrichtlichen Arbeit geben. Sie wollen einen wichtigen Beitrag zur Schul- und Unterrichtsentwicklung leisten. Lehrpersonen müssen informiert und geschult werden, ihren Unterricht stärker zu reflektieren und eigenverantwortlich weiter zu entwickeln. Die Bildungsstandards wollen der Autonomie einen Rahmen geben und durch das Setzen von Maßstäben die Verantwortlichkeit stärken. Den Lehrpersonen werden die Standards helfen, mit dem zunehmenden Rechtfertigungsdruck professionell umzugehen (vgl. Lucyshyn, 2006). Wenn Schulen aufgefordert werden, verstärkt Unterrichtsentwicklung zu betreiben, heißt das auch, sich regelmäßig des Erfolgs der Arbeit zu vergewissern (interne Evaluation) und sich einer „standardisierten“ Rückmeldung des Unterrichtserfolgs zu stellen (externe Evaluation). Bildungsstandards liefern hierfür die notwendigen Vergleichsmaßstäbe. Beim schulischen Lernen geht es um Wissen, um Haltungen, Einstellungen, Interessen und grundlegende Fertigkeiten, die Schüler erwerben sollen. In Lehrplänen werden diese in Lernzielen und -inhalten aufgelistet und zeitlich angeordnet. Bildungsstandards hingegen arbeiten die zentralen Kompetenzbereiche heraus, die im Laufe der schulischen Ausbildung aufgebaut werden sollen.

¹ Vgl. http://www.klassezukunft.at/statisch/zukunft/de/bildungsstandards_folder.pdf [26.05.2006]

² Vgl. http://www.klassezukunft.at/statisch/zukunft/de/arbeitsbericht_bildungsstandards_14_02_2006.pdf [26.05.2006]

³ Vgl. <http://www.gemeinsamlernen.at/index2.asp> [26.05.2006]

Die derzeitigen Bildungsstandards beschreiben die erwarteten Lernergebnisse auf der 4. oder 8. Schulstufe, Lehrpläne geben die Inhalte und Ziele vor. So wird deutlich, dass Bildungsstandards und Lehrpläne in einem inhaltlichen Zusammenhang stehen.

1.3. Wer entwickelt Bildungsstandards? Wo werden sie eingesetzt?

Zur Entwicklung von Bildungsstandards hat das österreichische Bildungsministerium Arbeitsgruppen, bestehend aus Fachdidaktikern und Schulpraktikern eingesetzt. Diese werden durch eine Steuergruppe unter wissenschaftlicher Beteiligung koordiniert. Bildungsstandards werden für die Nahtstellen der Bildungslaufbahn erarbeitet, und zwar für

- die 4. Schulstufe (Deutsch, Mathematik) und
- die 8. Schulstufe (AHS bzw. Hauptschule in Deutsch, Mathematik und Englisch).

Mit der Entwicklung und Überprüfung von Standard-Tests und der Testung in den Fächern Deutsch und Mathematik wurde die *Test- und Beratungsstelle* des Arbeitsbereiches Psychologische Diagnostik⁴ (Leitung: Univ. Prof. Dr. Mag. Klaus D. Kubinger) an der Fakultät für Psychologie der Universität Wien beauftragt. Für die Testentwicklung in Englisch erging der Auftrag an das *Language Testing Centre* an der Universität Klagenfurt⁵ (Leitung: ao. Univ. Prof. Dr. Günther Sigott). Wesentliche Belange der Organisation der Testungen und der Datenverarbeitung erfolgen am Pädagogischen Institut in Linz (Leitung: AL Mag. Wolfgang Schwarz) – siehe dazu auch das Organigramm im Anhang.

1.4. Wie sind die Bildungsstandards aufgebaut?

Die Abbildungen 1 und 2 sollen – beispielhaft für andere Bereiche – den Aufbau der Bildungsstandards Mathematik 8. Schulstufe (M8) skizzieren (Heugl, 2004; s. vor allem Bundesministerium für Bildung, Wissenschaft und Kunst, 2004). Weitere Erläuterungen sowie konkretere Informationen über Aufbau und Inhalt der Bildungsstandards sind auf der Website <http://www.gemeinsamlernen.at/index2.asp> [26.05.2006] beschrieben. Außerdem beinhaltet jede Testergebnis-Rückmeldung (vgl. in Kap. 3) das jeweilige Kompetenzmodell. Anzumerken ist, dass die Bildungsstandards jeweils auch durch sog. „Aufgabenbeispiele“ (s. unter der eben angeführten Website) illustriert sind; diese haben die Funktion, den Lehrern standardbezogenes Übungsmaterial für den Unterricht anzubieten (vgl. in Kap. 2, insbesondere den Unterschied zu den sog. „Testitems“).

⁴ www.testzentrum.at [26.05.2006]

⁵ www.uni-klu.ac.at/ltc [26.05.2006]

Kompetenzmodell Sekundarstufe I

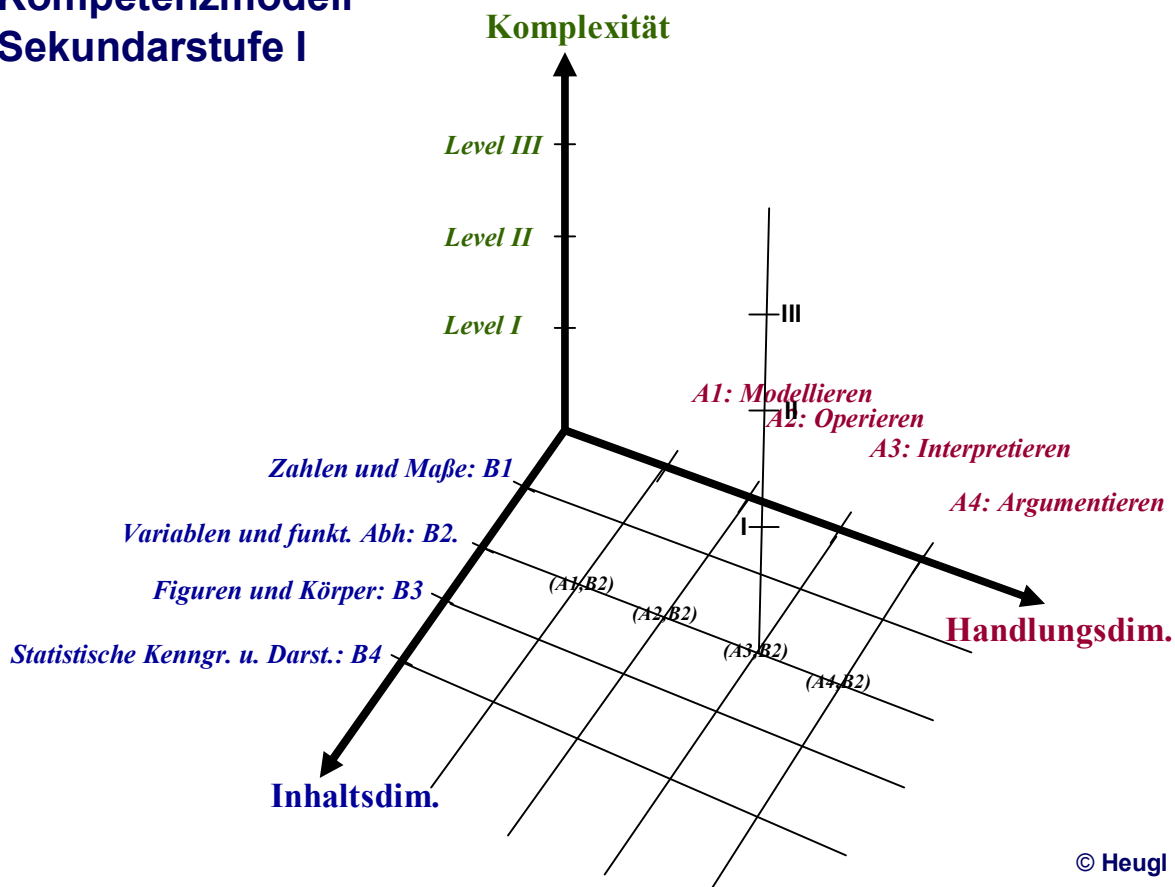


Abbildung 1: Kompetenzmodell Sekundarstufe I – Mathematik (Heugl, 2004; s. Bundesministerium für Bildung, Wissenschaft und Kunst, 2004)

Standard = Paar aus einem Element der Handlungsdimension und einem Element der Inhaltsdimension

Handlungsdimension (A)	Inhaltsdimension (B)
A4: Argumentieren Ich kann einzelne Rechenschritte begründen wie auch begründen, warum etwas falsch ist	B1: Arbeiten mit Zahlen und Maßen Ich kenne die Begriffe „Prozent“ und „Zinsen“ und kann damit verständlich umgehen

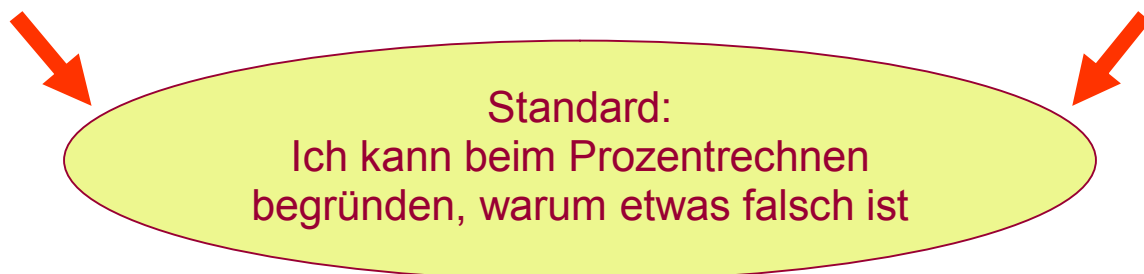


Abbildung 2: Standard-Beispiel: Kompetenzmodell Sekundarstufe I (Heugl, 2004; s. Bundesministerium für Bildung, Wissenschaft und Kunst, 2004)

Die österreichischen Bildungsstandards werden in Broschüren generell wie folgt festgelegt:

- Eine Präambel definiert den Beitrag des jeweiligen Faches zur Bildung.
- Die Kompetenzbereiche des jeweiligen Faches sind beschrieben, die Standards formuliert.
- Aufgabenbeispiele in unterschiedlichen Komplexitätsstufen illustrieren die Standards.

Die Kompetenzmodelle sind für alle betroffenen Unterrichtsfächer auf der Website <http://www.gemeinsamlernen.at/index2.asp> [26.05.2006] beschrieben.

1.5. Ab wann gelten die Bildungsstandards? Wie wird das Erreichen der Standards überprüft?⁶

Die Erprobungsphase läuft bis 2007, dann wird das Ergebnis der Bundesregierung zur Beschlussfassung vorgelegt. Hinsichtlich der Implementierung der Bildungsstandards sind in allen Bundesländern Informationsveranstaltungen für die Schulaufsichtsbehörde, die Schulleitung und die Pädagogischen Institute vorgesehen. Einen besonderen Schwerpunkt stellt in diesem Zusammenhang die Lehreraus- und -fortbildung dar. In den einzelnen Bundesländern werden unter Mitwirkung der Arbeitsgemeinschaften Fortbildungsangebote für die Lehrer vorbereitet.

Die Überprüfung der Standards erfolgt je Unterrichtsgegenstand durch einen Test. Während der Pilotphase werden erste Erfahrungen mit der Testung gesammelt, mit der Kalibrierung der einzelnen Tests wird begonnen. Die erste österreichweite Überprüfung wird nach erfolgreichem Abschluss der Pilotphase im Schuljahr 2007/2008 stattfinden. Pro Jahr werden 30 % der 4. und 8. Schulstufen überprüft (zum gegebenen Zeitpunkt ist vorgesehen: Jeweils 10 % in Deutsch, Mathematik und Englisch, d.h. pro Schule wird voraussichtlich nur jeweils ein Fach getestet).

1.6. Was geschieht mit den Ergebnissen?

Die mit dem Test gewonnenen Ergebnisse werden Schülern (individuell, datengeschützt), Lehrern (für die betreffende Klasse bzw. Gruppe, datengeschützt), der Schulleitung (für die betreffende Schule, datengeschützt) und der Schulverwaltung (für das Bundesland) rückgemeldet, aber nicht veröffentlicht. Keine Ebene der Beteiligten erhält Ergebnisse aus einer darunter liegenden Ebene.

Die involvierten Zielgruppen (Schüler, Lehrpersonal, Schulleitungen, Schulverwaltung) erhalten die entsprechende Ergebnismeldung der Ergebnisse über einen speziellen Zugangscode einer Plattform im Internet, und zwar auf der Seite www.bildung-standards.at [26.05.2006]. Dort kann jeder Schüler seine persönlichen Ergebnisse abfragen und sieht im Vergleich dazu das Ergebnis seiner Klasse bzw. seiner [Leistungs-] Gruppe (s. Kap. 3). Der jeweilige Lehrer kann das bezogen auf die einzelnen Schüler anonymisierte Ergebnis seiner Klasse bzw. Gruppe abfragen und dieses mit anderen Klassen desselben Leistungsgruppenniveaus (1., 2., 3., Leistungsgruppe, AHS bei der 8. Schulstufe bzw. 4. Schulstufe) in Österreich vergleichen. Die Schulleitung bekommt das Ergebnis ihrer Schule im Vergleich zum gesamtösterreichischen Ergebnis. Die jeweilige Schulaufsichtsbehörde erhält das Ergebnis des betreffenden Bundeslandes im Vergleich zum Ergebnis ganz Österreichs.

⁶ Vgl. http://www.klassezukunft.at/statisch/zukunft/de/bildungsstandards_folder.pdf [26.05.2006]

Ein Ranking etwa von Schulen oder Bundesländern wird also explizit ausgeschlossen, weil jeweils nur die betroffene Ebene die für sie relevanten Daten erhält. Die allgemeine Öffentlichkeit erhält keine Testergebnisse. Das genaue Procedere der Rückmeldung ist weiter unten beschrieben.

1.7. Vergleich von PISA und Bildungsstandards (vgl. Lucyshyn, 2006)

Die Gemeinsamkeiten der PISA-Studie und der Bildungsstandards sind:

- Beides sind Regelstandards, welche erwartete Schülerkompetenzen an bestimmten Stellen des Bildungsweges ausdrücken.
- Z.B. auf Mathematik bezogen, erfassen beide die mathematische Grundbildung: Von der Rechenfertigungsorientierung bis zur Problemlöse-Orientierung.
- Beide arbeiten mit Tests, die unterschiedliche Antwortformate verwenden (vgl. dazu weiter unten).
- Bei beiden handelt es sich um Instrumente zur Systemevaluation (sog. „Monitoring“).
- Bei beiden ist die statistisch-testtheoretische Grundlage das sog. „Rasch-Modell“ (vgl. in Kap. 2 und im Anhang).

Unterschiede bestehen in folgenden Aspekten:

- Bildungstheoretischer Orientierungsrahmen
 - o PISA: Der Schwerpunkt liegt auf der funktionalen Anwendung von Kenntnissen in ganz unterschiedlichen Kontexten. Es wird nicht versucht, Lehrpläne beteiligter Länder zu berücksichtigen (vgl. das PISA Framework in Tab. 1).
 - o Österreichische Bildungsstandards: Grundlage ist der Lehrplan mit der Bildungs- und Lehraufgabe und den vorgeschriebenen Inhaltsbereichen sowie das jeweilige Kompetenzmodell (vgl. ebenfalls Tab. 1).
- Vergleichbarkeit
 - o PISA: internationale Vergleichbarkeit
 - o Österreichische Bildungsstandards: Vergleichbarkeit innerhalb des österreichischen Bildungswesens in Bezug auf verschiedene Aspekte (schulartenspezifisch – VS, AHS, HS – und bei letzterer innerhalb der Leistungsgruppen; regional, bundesländerweise; schulweise; lehrerweise; schülerweise).
- Ergebnissrückmeldung
 - o Bei den österreichischen Bildungsstandards ist im Gegensatz zu PISA eine Begleitung vor allem für die Schulen vorgesehen. Auch die Schüler erhalten eine Ergebnissrückmeldung, anhand derer die Stärken und Schwächen detailliert dargestellt werden; so kann ein entsprechender Förderbedarf aufgezeigt werden. Demzufolge sind auch spezielle Fortbildungs- und Unterstützungsangebote für Lehrer in Vorbereitung.

Die Unterschiede zwischen Kompetenzmodell der Bildungsstandards und dem PISA Framework werden beispielhaft für Mathematik in Tabelle 1 dargestellt.

Tabelle 1: Vergleich des Kompetenzmodells „Sekundarstufe I – Mathematik“ der österreichischen Bildungsstandards mit dem PISA Framework (nach Lucyshyn, 2006).

Kompetenzmodell - Bildungsstandards	PISA Framework
<p>Handlungsdimension:</p> <ul style="list-style-type: none"> • Modellbilden, Darstellen • Operieren, Rechnen • Interpretieren und Dokumentieren • Argumentieren und Begründen <p>Inhaltsdimension:</p> <ul style="list-style-type: none"> • Arbeiten mit Zahlen und Maßen • Arbeiten mit Variablen und funktionalen Abhängigkeiten • Arbeiten mit Figuren und Körpern • Arbeiten mit statistischen Kenngrößen und Darstellungen <p>Komplexitätsdimension: 3 Stufen gemäß „Kognitiver Komplexität“: Diese erfasst Anforderungen an Ausmaß, Intensität, und Vielschichtigkeit von Denkvorgängen beim Lösen von Aufgaben</p> <p>Kontext: Nicht explizit Bestandteil des Kompetenzmodells, wird aber bei der Entwicklung der Standard-Tests sehr stark beachtet.</p>	<p>Performance dimension:</p> <ul style="list-style-type: none"> • Denken und Schlussfolgern • Argumentieren • Kommunizieren • Modellbildung • Probleme formulieren und lösen • Repräsentieren • Anwenden symbolischer, formaler und technischer Sprache • Anwendung von Hilfsmitteln <p>Content dimension:</p> <ul style="list-style-type: none"> • Raum und Form • Veränderungen und Zusammenhänge • Größen • Unsicherheit <p>Leistungsstufen: 7 Stufen gemäß den testtheoretisch empirisch abgeleiteten Lösungsschwierigkeiten der Tests</p> <p>Context:</p> <ul style="list-style-type: none"> • Persönliches Umfeld • Öffentliches Umfeld • Wissenschaftliches Umfeld • Umfeld Schule/Beruf

2. Prinzipien bei der Erstellung der Standard-Tests

Alles im Folgenden genauer Ausgeführte ist der einschlägigen wissenschaftlichen Literatur zu entnehmen (vgl. am besten dazu Kubinger, 2006).

Im Gegensatz zu den unter 1.4. angesprochenen „Aufgabenbeispielen“, die den Standards als didaktisches Mittel unmittelbar selbst angehören, wird im Folgenden von „Testitems“⁷ (kurz: Items) gesprochen, wenn es sich um die einzelnen Prüfaufgaben in den verschiedenen Standard-Tests handelt.

2.1. Testtheoretische Prinzipien bei der Itemkonstruktion zu den Unterrichtsfächern Mathematik und Deutsch⁸

Grundsätzlich müssen Testitems dem Anspruch der psychologischen Testtheorie genügen. Im vorliegenden Standardprojekt kommt dafür das sog. „Rasch-Modell“ (vgl. die Zusammenstellung im Anhang) zur Anwendung. Außerdem müssen Testitems nach einschlägigen psychologischen Gestaltungsregeln konzipiert sein. Konkret liegen den Standard-Tests folgende Prinzipien zugrunde:

- Items werden nur hinsichtlich gelöst oder nicht gelöst bewertet, d.h. es gibt keine irgendwie gewichteten Gutpunkte für in bestimmter Weise teilrichtige Antworten. Die Gründe dafür sind, dass 1) eine faire Verrechnung von Testleistungen zu Testwerten gemäß Testtheorie schon aufwendig genug ist, wenn es sich um eine einfache zweikategorielle Verrechnung handelt – eine faire mehrkategorielle Verrechnung ist unverhältnismäßig schwieriger zu gewährleisten; dass 2) die Auswertung eindeutig und damit objektiv erfolgt.
- Items dürfen nicht aufeinander aufbauen, d.h., die Lösung jedes einzelnen Items darf nicht davon abhängig sein, ob irgendein vorausgehendes Item gelöst oder nicht gelöst worden ist. Insbesondere soll zu einem Thema (zu einer inhaltlichen Angabe) nur ein einziges Item formuliert werden bzw. soll immer nur ein einziges Item zur selben inhaltlichen Angabe ein und demselben Schüler vorgegeben werden. Der Grund dafür liegt in der andernfalls gegebenen Übergewichtung eines bestimmten Themas (einer inhaltlichen Angabe), was zu unfairer Verrechnung bei solchen Schülern führen kann, die zufällig bei diesem Thema gehandicapt sind (Missverständnis; Desinteresse; Demotivation u.a.).
- Items dürfen lediglich die zu messen gesuchte Fähigkeit erfassen, d.h. zum Beispiel Faktoren, wie 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 taxativ bekannt sein. Letzteres ist v.a. für Items mit freiem Antwortformat relevant.

⁷ Für Aufgaben, Fragen bzw. *Statements* wird in der Psychologie der aus dem Englischen kommende Oberbegriff „Item“ verwendet.

⁸ Das teilweise davon abweichende Vorgehen für das Unterrichtsfach Englisch s. in einem später erscheinendem Supplement.

- Die sprachliche Formulierung der Items muss extrem einfach sein, es darf nicht von der Sprachkompetenz (bei Mathematik) bzw. von (Fremd-) Wortschatz (bei Deutsch) u.ä. abhängen, wie schwer die Lösung eines Items einem Schüler fällt.

Davon abgesehen besteht das Prinzip, dass nicht allen Schülern (auch desselben Leistungsgruppenniveaus) dieselben Items vorgegeben werden; dies liegt daran, dass

- es als Testhefte mehrere Parallelförmungen zur Verhinderung von Abschreiben benachbart sitzender Schüler geben muss (in der Erprobungsphase sind es teilweise nur jeweils zwei Testformen pro Klasse bzw. Gruppe);
- ein besonders großer Itempool eingesetzt werden soll, um die Standards umfassend zu prüfen. Es handelt sich im Wesentlichen um ein Monitoring, sodass es nicht darauf ankommt, die Leistungen einzelner Schüler möglichst genau, jedoch über die gesamte Schülerschaft in Bezug auf die Inhalte repräsentativ zu messen. Daher sind am besten etwa 500 Items je Standard zu verwenden, um Zufallseinflüsse bei der Itemzusammenstellung (vor allem thematisch oder formulierungsmäßig bedingt) zu minimieren. Die Vergleichbarkeit der Testleistungen von Schülern, die unterschiedliche Items bearbeitet haben, ist trotzdem über das Rasch-Modell gewährleistet.
- Dann geht es um die Maximierung der „Halbwertszeit“ (das ist die Zeit bis zum Bekanntwerden einzelner Items in der Hälfte der betreffenden Population); je mehr Items gleichzeitig eingesetzt werden, umso weniger kann ein einzelnes Item samt Lösung bekannt werden und damit die Ergebnisse artifiziell beeinflussen, umso weniger verleitet es Betroffene, sich um das Bekanntwerden von Items zu bemühen, bzw. die Öffentlichkeit, an einzelne Items heranzukommen.

Weiters müssen jährlich neue Items ergänzend entwickelt werden; dies hat den Zweck, dass

- die Laufzeit der aktuell eingesetzten Items möglichst niedrig ist, um das Bekanntwerden von Items möglichst gering zu halten;
- die Akzeptanz seitens der Öffentlichkeit zu erhöhen, indem der Itempool eben so viele Facetten des Standards (auch inhaltlich bezogen) wie möglich abdeckt.

Schließlich sind je nach Leistungsgruppenniveau unterschiedlich schwierige Items in die jeweiligen Testformen aufzunehmen, um dem Leistungsgrad der Schüler zu entsprechen – trotzdem ist über das Rasch-Modell ein Vergleich der erbrachten Leistungen möglich.

2.2. Antwortformate

Es werden verschiedene Antwortformate eingesetzt:

- freies Antwortformat mit freiem Text: Die Antwort in Form eines freien Textes wird von Fachleuten auf Richtigkeit bewertet – aus Gründen der Ökonomie kommen solche Items relativ selten vor, jedoch bei jedem Schüler mindestens einmal.

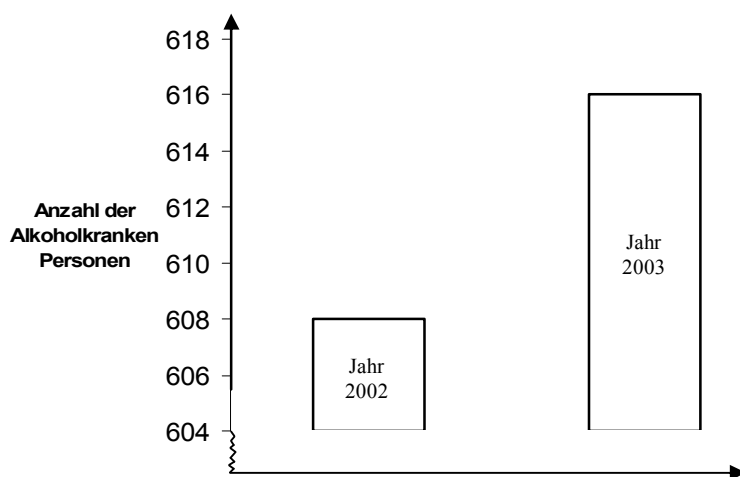
- freies Antwortformat im Kästchenformat (nur bei Mathematik): Die (alpha-) numerische Lösung muss von dem Schüler in entsprechende (Stellenwert-) Kästchen eingetragen werden; die Anzahl der Kästchen entspricht immer genau der benötigten Anzahl.
- Multiple-Choice Antwortformat, Variante „1 aus 6“: Es sind 6 Antwortmöglichkeiten vorgegeben, darunter die Lösung sowie 5 Distraktoren (d.s. Antwortvorschläge, die der Lösung nahe kommen, aber eben in gewisser Weise falsch sind). Die Schüler werden jedes Mal darüber informiert, dass genau eine einzige Antwort richtig ist.
- Multiple-Choice Antwortformat, Variante „2 aus 5“: Es sind 5 Antwortmöglichkeiten vorgegeben, darunter befinden sich genau 2 richtige Antworten sowie 3 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 die Aufgabe als richtig verrechnet wird.
- Multiple-Choice Antwortformat, Variante „2 x (2 aus 3)“ (nur bei Deutsch): In einem Lückentext fehlen zwei Textstellen, für die je 3 Antwortmöglichkeiten vorgegeben werden. Die Schüler werden jedes Mal darüber informiert, dass sie für beide Lücken die richtige Antwort ankreuzen müssen, damit die Aufgabe als richtig verrechnet wird.

Im Folgenden werden Beispiele zu den Antwortformaten aus dem Mathematik-Standard-Test für die 8. Schulstufe (M8) gegeben.

2.2.1. Freies Antwortformat mit freiem Text

Dieses Antwortformat wird hauptsächlich dafür genutzt, die Kompetenz des Argumentierens und Begründens (M8) zu prüfen.

In einer Zeitschrift ist zu lesen: „Untenstehende Graphik demonstriert, dass die Anzahl der Alkoholkranken in der Stadt X von 2002 bis 2003 stark zugenommen hat“ Ist diese Aussage gerechtfertigt? *Begründe im Antwortbogen!*



(Lösung: „Die Aussage ist nicht gerechtfertigt, weil 8 Alkoholranke mehr sind keine **STARKE** Zunahme bei rund 600.“ Oder: „Eine Zunahme von 608 auf 616 ist keine **STARKE** Zunahme.“ Bzw. über die Berechnung des Prozentsatzes der Veränderung.)

2.2.2. Freies Antwortformat im Kästchenformat

Bei einer Spendensammlung für bedürftige Kinder wurden in 9 Klassen einer Schule folgende Beträge gesammelt:

1a 29 €	2b 56 €	3b 39 €
1b 35 €	2c 45 €	4a 58 €
2a 27 €	3a 37 €	4b 34 €

Wie viel Geld wurde insgesamt gesammelt?

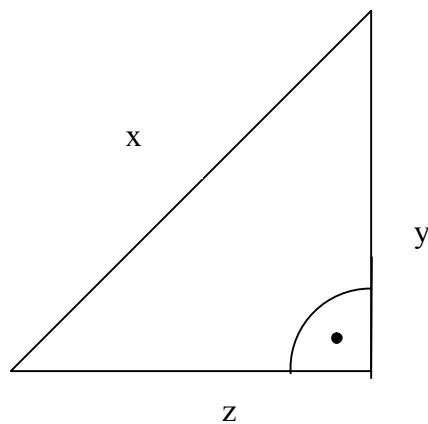
Trag die berechnete Lösung im Antwortbogen ein!

Lösung: €

2.2.3. Multiple-Choice „1 aus 6“

Beim Multiple-Choice Antwortformat ist zu bedenken, dass grundsätzlich die Möglichkeit besteht, die richtige Antwort zu erraten. Gäbe es nur 3 Antwortmöglichkeiten, von denen eine richtig ist, so beträgt die Wahrscheinlichkeit, die richtige Antwort zu erraten, ohne auch nur die geringste Kompetenz zu haben, 0,333, das entspricht 33,3 %. Bei der hier vorgesehenen Version beläuft sich dagegen diese „Ratewahrscheinlichkeit“ auf 1/6, das entspricht 16,7 %.

Welche der angegebenen Formeln ermöglicht die richtige Berechnung der Länge der Seite y des abgebildeten rechtwinkligen Dreiecks? Markiere den entsprechenden Buchstaben im Antwortbogen!



- A $y = x^2 - z^2$
- B $y = \sqrt{x^2 + z^2}$
- C $y = \sqrt{x^2 - z^2}$
- D $y = \sqrt{z^2 - x^2}$
- E $y = \sqrt{x - z}$
- F $y = x - z$

(Lösung: C)

2.2.4. Multiple-Choice „2 aus 5“

Vielfach ist es auch möglich, zwei verschiedene richtige Antworten festzulegen. Die Ratewahrscheinlichkeit für Schüler ohne jede Kompetenz beträgt dabei nur 0,10, das entspricht 10%.

Peter besitzt 100 €, Josef besitzt 200 €. Josef sagt stolz zu Peter: „Ich besitze um 100 % mehr Geld als du.“ Peter entgegnet: „Das macht mir gar nichts aus, ich habe ja nur um 50 % weniger als du.“ Zwei der folgenden fünf Behauptungen klären das Ganze auf. Finde aus den Antwortmöglichkeiten A bis E diese beiden Sätze. *Markiere die entsprechenden Buchstaben im Antwortbogen!*

- A Irgendwo ist hier ein Widerspruch. 50 % können nicht 100 % sein.
- B Peter täuscht sich. Er besitzt 100 % weniger als Josef.
- C Beide haben Recht, weil sie von verschiedenen Grundwerten ausgehen
- D Peter hat Recht: 50 % von 200 € sind 100 €.
- E Josef täuscht sich, weil er um 200 % mehr besitzt als Peter.

(Lösung: C, D)

3. Ergebnismeldung

Die Ergebnisse werden, wie in Abschnitt 1.6 ausgeführt, getrennt für Schüler, Lehrer, Schulleiter und Schulaufsichtsbehörde rückgemeldet. Die Ergebnismeldungen sind strukturell ähnlich aufgebaut.

3.1. Ergebnismeldung auf Schülerebene

Es werden die Ergebnisse eines Schülers sowohl bezogen auf den Gesamttest als auch bezogen auf die einzelnen Kompetenzbereiche rückgemeldet (beim Englisch-Standard-Test für die 8. Schulstufe, E8, besteht das Gesamtergebnis aus zwei Teilen, der Leistung im Leseverständnis und der Leistung im Hörverständnis). Auf der Übersichtsseite erhalten die Schüler zunächst eine Beschreibung der einzelnen Kompetenzbereiche. Diese Beschreibung weicht von jener ab, die in den Rückmeldungen auf den anderen Ebenen zu finden ist. Grund dafür ist eine Vereinfachung dieser Beschreibungen, damit sie die Schüler besser verstehen können. Für die Betroffenen der anderen Ebenen ist es möglich, die für die Schüler vereinfachten Beschreibungen der einzelnen Kompetenzbereiche ebenfalls einzusehen.

3.1.1. Generell

Das erste Ergebnis besteht aus der Information darüber, wie viele Items (dort bezeichnet als „Testaufgaben“) insgesamt im Testheft enthalten waren, wie viele davon der jeweilige Schüler bearbeitet hat und wie viele davon richtig beantwortet (gelöst) wurden. Erfahrungen haben gezeigt, dass die Zeit ausreichend ist, sodass die meisten Schüler alle Items bearbeiten können. Dennoch werden manchmal Items ausgelassen.

Die Ergebnismeldungen zu den einzelnen Kompetenzbereichen erfolgen dort, wo dies aufgrund der testtheoretischen Analysen möglich ist, über so genannte Prozentränge. Ein Prozentrang von beispielsweise 40 drückt aus, dass 40 % der „Vergleichsstichprobe“ schlechter oder gleich

gut abgeschnitten haben. Als „Vergleichsstichprobe“ werden alle getesteten Schüler desselben Leistungsgruppenniveaus (4. Schulstufe bzw. bei 8. Schulstufe: 1., 2. oder 3. Leistungsgruppe bei Hauptschulen oder Kooperativen Mittelschulen sowie AHS) herangezogen⁹. Das bedeutet, dass ein Schüler, der sich zum Testzeitpunkt in der 1. Leistungsgruppe befand, mit allen anderen getesteten Schülern aller 1. Leistungsgruppen verglichen wird. Der Prozentrang wird für die Schüler grafisch veranschaulicht (Abb. 3):

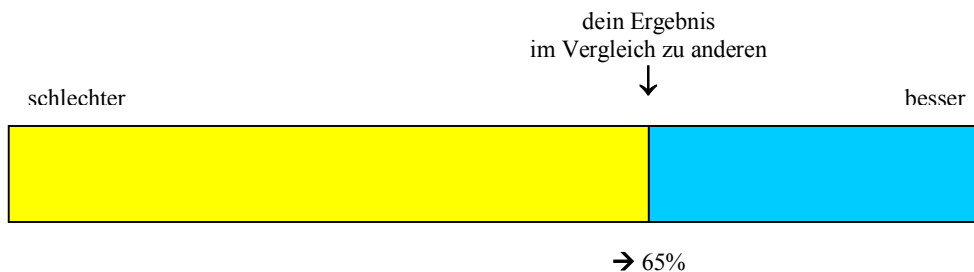


Abbildung 3: Grafische Veranschaulichung des Prozentranges für Schüler

Sofern testtheoretische Analysen zeigten, dass ein solcher Vergleich nicht fair möglich ist, wird dem Schüler nur rückgemeldet, wie viele Items eines bestimmten Kompetenzbereiches sie bzw. er bearbeitet hat und wie viele sie bzw. er gelöst hat.

Je Kompetenzbereich hat der Schüler die Möglichkeit, sich an Hand von typischen Beispielen von Testitems zu veranschaulichen, was genau dabei geprüft wurde – es handelt sich um Items, die in Zukunft nicht mehr im Standardtest eingesetzt werden. Zusätzlich kann er sich an Hand didaktisch ausgearbeiteter Übungsbeispiele informieren, wie Aufgaben zu diesem Kompetenzbereich im Detail zu lösen sind.

Am Ende erhalten die Schüler noch generelle Lerntipps. Diese sind als Fragen gestaltet, die ein Reflektieren der eigenen Lernstrategie auslösen sollen. Diese Lerntipps sollen mittelfristig noch ergänzt werden; die derzeitige Fassung (in Anlehnung an Mandl und Friedrich, 2006 sowie Konrad, 1999) sieht Folgendes vor – Punkt 2) beispielhaft für Mathematik¹⁰:

1) Fragen und Tipps zum Thema „Lernen“ – allgemein:

- Beginnst du rechtzeitig zu lernen (am besten ungefähr zwei Wochen vor einer Prüfung oder einer Schularbeit)? Dann gerätest du kurz vor der Prüfung nicht in Zeitnot, sondern hast Zeit zum Wiederholen und kannst so dein Lernpensum besser bewältigen.
- Teilst du dir den Lernstoff in „machbare“ Einheiten ein, so dass du an jedem Tag „ein bisschen“ lernst? Es ist besser, jeden Tag eine Stunde zu lernen, als an einem Tag fünf Stunden am Stück.
- Teilst du dir den Stoff logisch ein? Es ist leichter, z.B. jeden Tag ein inhaltlich abgeschlossenes Kapitel zu lernen.
- Planst du auch abschließende Tage ein, an denen du nur mehr wiederholst? So kannst du das Gelernte festigen und musst nicht am Ende noch etwas Neues lernen.

⁹ Für die Rückmeldung zur Standard-Testung zu M8 im Jahr 2006 wurde eine Trennung zwischen Schülern der 1. Leistungsgruppe und Schülern der AHS noch nicht vorgenommen, diese erfolgt ab 2007.

¹⁰ Vgl. http://fb04130.mathematik.tu-darmstadt.de/mathezirkel/index.php?page_id=14&php [26.05.2006]

- Machst du regelmäßig Pausen? Nach 30 Minuten Lernen sollte man fünf Minuten Pause machen. Du kannst dich danach wieder besser konzentrieren.
- Wenn du viele Tage hintereinander lernst, planst du dann auch einmal einen Tag Pause ein? Dann bist du erholter und kannst mit neuer Energie weitermachen.
- Um welche Uhrzeit lernst du? Die beste Lernzeit ist am Vormittag und am späten Nachmittag/frühen Abend.
- Nachdem du etwas geschafft hast, belohnst du dich mit angenehmen Tätigkeiten? Wenn du dich auf etwas freuen kannst, fällt dir das Lernen vielleicht leichter.
- Hast du genügend Platz auf deinem Schreibtisch? Hältst du deinen Schreibtisch sauber und ordentlich? Dann bist du weniger abgelenkt.
- Hast du genügend Licht an deinem Schreibtisch? Dann ist das Lernen für die Augen nicht so anstrengend.

2) Fragen und Tipps zum Thema „Üben in Mathematik“:

Beim Lernen zu Hause:

- Übung macht den Meister, gerade in Mathematik. Übe die Beispiele, die in der Schule durchgenommen wurden, zu Hause nochmals.
- Liest du dir die Aufgabe sorgfältig Satz für Satz durch? Du vermeidest damit, dass du etwas falsch verstehst oder etwas übersiehst.
- Markierst du dir, welche Angaben du bereits bearbeitet hast und welche dir noch fehlen? Du siehst dann leichter, was du zu tun hast.
- Veranschaulichst du dir die Aufgabe? Machst du dir eine Skizze? Du gewinnst so einen besseren Überblick über die Aufgabenstellung.
- Versuchst du die Aufgabe mit eigenen Worten neu zu formulieren? Denkst du dir selbst eine ähnliche Aufgabe aus? Wenn du dir ein praktisches Beispiel überlegst, wird dir vielleicht klarer, wie man eine Aufgabe lösen kann.
- Überlegst du dir am Ende, ob das Ergebnis bzw. die Lösung sinnvoll ist? Damit kannst du Rechenfehler entdecken.
- Versuchst du anderen Mitschülern die Lösung einer Aufgabe zu erklären? Das hilft dir auch selbst, die Aufgabe besser zu verstehen.

In der Schule:

- Schreibst du dir in der Schule die Erklärungen deines Lehrers mit? Wenn du zu Hause ähnliche Aufgaben übst, können dir diese Erklärungen helfen.

Bei einer Prüfung oder einer Schularbeit:

- Liest du dir die Aufgabe sorgfältig Satz für Satz durch? Du vermeidest damit, dass du etwas falsch verstehst oder etwas übersiehst.
- Veranschaulichst du dir die Aufgabe? Machst du dir eine Skizze? Du gewinnst so einen besseren Überblick über die Aufgabenstellung.
- Überlegst du dir am Ende, ob das Ergebnis bzw. die Lösung sinnvoll ist? Damit kannst du Rechenfehler entdecken.

3.1.2. Vierte Schulstufe

Bei Kindern der 4. Schulstufe wird aus Gründen der besseren Verständlichkeit auf die Angabe der Prozentränge verzichtet; sie bekommen die Anzahl bearbeiteter bzw. gelöster Items rückgemeldet. Die einzelnen Kompetenzbereiche werden nach der individuell erbrachten Leistung geordnet (absteigend, d.h. die beste Leistung zuerst usw.) und in Relation zur Vergleichsgruppe grafisch ver-

anschaulicht. Die Schüler erhalten die Erklärung: „Der gelbe Balken zeigt dir deine Testleistung. Der gelbe Balken ist kurz: das bedeutet, du hast wenige Aufgaben gelöst. Der gelbe Balken ist lang: das bedeutet, du hast viele Aufgaben gelöst.“ (vgl. Abb. 4, beispielhaft aus der Schülerrückmeldung zu D4). Grundlage für den Vergleich sind die aus den Fähigkeitsparametern (bei Geltung des Rasch-Modells) bzw. aus der Anzahl gelöster Aufgaben errechneten Prozentränge.

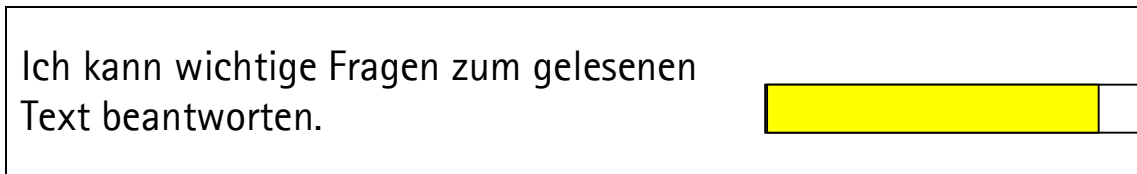


Abbildung 4: Beispiel zur Schülerrückmeldung D4

3.2. Ergebnissrückmeldung auf Lehrerebene

Lehrer erhalten für jede ihrer getesteten Klassen bzw. [Leistungs-] Gruppen einen eigenen Login-Code. Hat ein Lehrer beispielsweise eine 1. Leistungsgruppe und eine 3. Leistungsgruppe so erhält er zwei Login-Codes, einen für die 1., einen für die 3. Leistungsgruppe (selbst dann, wenn er zwei 1. Leistungsgruppen verschiedener Klassen unterrichtet). Im Gegensatz zur Schülerebene werden den Lehrern auch ausführliche Informationen zum Hintergrund der Testungen gegeben.

Die Übersichtsseite ist ähnlich aufgebaut, wie jene der Schülerebene. Zusätzlich zum Gesamtergebnis und den Ergebnissen zu den einzelnen Bereichen wird angegeben, ob sich die Zusammensetzung in der jeweiligen Klasse bzw. Gruppe von der durchschnittlichen Zusammensetzung anderer (vergleichbarer) Klassen bzw. Gruppen unterscheidet: Die Schüler geben im Rahmen der Testung die demografischen Variablen Alter, Geschlecht und Muttersprache an. In den testtheoretischen Analysen wird überprüft, ob diese Variablen einen Einfluss auf das Niveau des Testergebnisses haben. Ist dies nicht der Fall, müssen die Testergebnisse selbst im Fall gegenüber sonst in Österreich abweichender Gruppenzusammensetzung bei der Interpretation nicht relativiert werden. Zeigen sich allerdings derartige Einflüsse und ist zudem eine deutlich abweichende Gruppenzusammensetzung gegeben, so werden diese in der Ergebnissrückmeldung dargestellt und sollten daher bei der Interpretation Berücksichtigung finden. Dies ist wie folgt vorgesehen (beispielhaft für Mathematik, einen Lehrer mit 1. Leistungsgruppe):

Zum Beispiel für Lehrer, bei denen die Verteilung in Bezug auf Kinder mit Deutsch als Muttersprache und Kinder mit einer anderen Muttersprache als Deutsch bedeutend vom Durchschnitt aller getesteter Schulen in Österreich abweicht – „bedeutend“ ist eine Abweichung von mindestens 10 % –, heißt es:

„Bei den Ergebnissen Ihrer 1. Leistungsgruppe gilt es folgendes zu berücksichtigen: Der Anteil an Schüler/innen mit anderer Muttersprache als Deutsch ist höher als dies in anderen 1. Leistungsgruppen bzw. in allen anderen Schulklassen in Österreich durchschnittlich der Fall ist.

Das ist insofern relevant, als Schüler/innen mit anderer Muttersprache als Deutsch im Standard-Test durchschnittlich um ca. * Testaufgaben weniger lösen als Schüler mit deutscher Muttersprache. Aus förderungsorientierter Sicht heißt das, dass Schüler mit anderer

Muttersprache als Deutsch besonderen Nachholbedarf haben betreffs der Entwicklung ihrer mathematischen Kompetenzen. Gezielte Förderungsmaßnahmen – auch in Bezug auf die deutsche Sprache – werden hier empfohlen.“

Sodann werden den Lehrern wesentliche Informationen zur Testung mitgeteilt (am Beispiel einer 1. Leistungsgruppe – „**“ bedeutet, dass diese Information je Standardtestung verschieden ist):

Die Schüler/innen bearbeiteten in einer **-minütigen Testzeit ** Testaufgaben. Für das Verständnis der einzelnen Ergebnisse ist folgendes wichtig zu wissen:

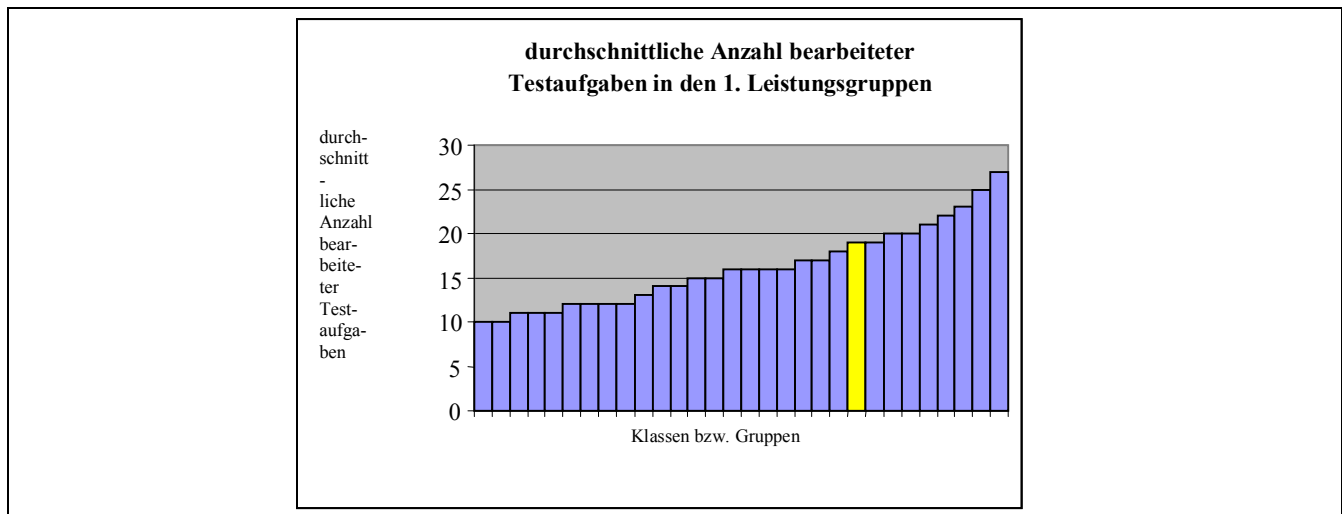
1) Es werden nicht alle ** Testaufgaben in die Ergebnisse einbezogen, sondern nur jene, die nicht mehr in der testtheoretischen Erprobung stehen. Es kommt daher vor, dass eine geringere Anzahl als ** bei den Ergebnissen als „vorgegeben“ angeführt wird; diese geringere Anzahl bezieht sich also auf die tatsächlich verrechneten Testaufgaben (Informationen zu den testtheoretischen Analysen erhalten Sie im Handbuch).

2) Den Schüler/innen Ihrer 1. Leistungsgruppe wurden ** unterschiedliche Testformen vorgegeben. Für diese Testformen werden u.U. unterschiedlich viele Testaufgaben berücksichtigt. Die Abweichungen sind jedoch geringfügig. Die Aussage über Ihre Klasse (Gruppe) ist daher über die Testformen gemittelt.

Das erste Ergebnis besteht aus Informationen über die durchschnittliche Anzahl vorgegebener, bearbeiteter und gelöster Items, wobei die durchschnittliche Bearbeitungs- und Lösungshäufigkeit der eigenen Klasse (bzw. Gruppe) mit jenen der vergleichbaren Gruppen in Bezug gesetzt werden. Ein Lehrer einer 1. Leistungsgruppe, zum Beispiel, erfährt also, wie seine Gruppe im Vergleich zu allen getesteten Schülern aller 1. Leistungsgruppen in Österreich abgeschnitten hat. Zur Veranschaulichung werden verschiedene Abbildungen dargeboten:

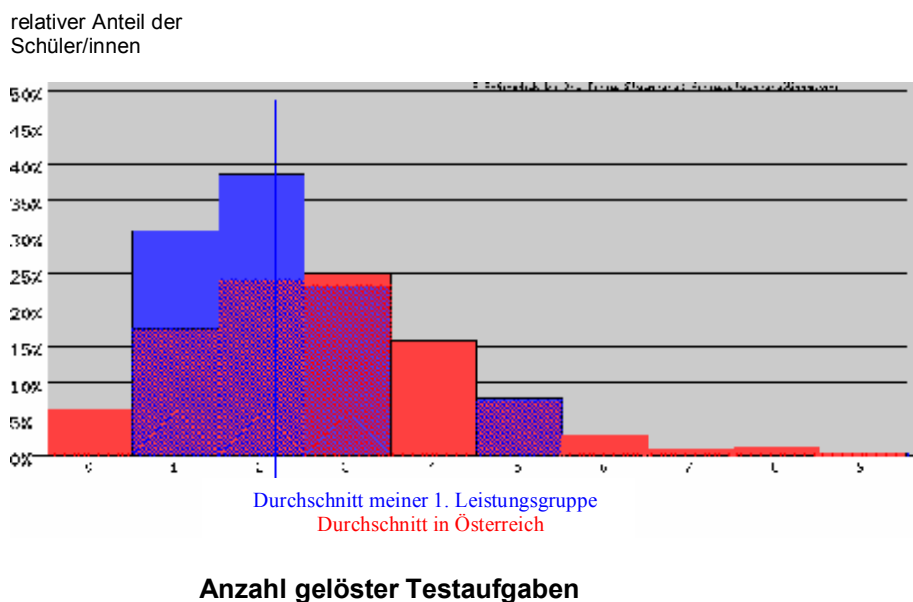
Abbildungstyp 1 – Variante „Gesamt-bearbeitet“: Vergleich der durchschnittlichen Anzahl bearbeiteter Testaufgaben mit allen anderen 1. Leistungsgruppen in Österreich

Sie sehen hier, wie viele Testaufgaben die Schüler/innen in Ihrer 1. Leistungsgruppe im Vergleich zu allen anderen 1. Leistungsgruppen in Österreich durchschnittlich bearbeitet haben. Jede Säule repräsentiert eine bestimmte 1. Leistungsgruppe in Österreich. Ihre 1. Leistungsgruppe ist andersfarblich hervorgehoben.



Abbildungstyp 2 – Variante „Gesamt-bearbeitet“: Die genauen Testleistungen Ihrer 1. Leistungsgruppe – Anzahl bearbeiteter Testaufgaben

Die blauen Säulen in der Grafik veranschaulichen den relativen Anteil der Schüler/innen in Ihrer 1. Leistungsgruppe, der eine bestimmte Anzahl von Testaufgaben bearbeitet hat. Die roten Säulen drücken aus, wie diese relativen Bearbeitungshäufigkeiten über alle Schüler/innen der 1. Leistungsgruppen in ganz Österreich verteilt sind. Im Fall der Überlappung sind die Säulen blau-rot schraffiert.



Diese beiden Grafiken beziehen sich auf die Anzahl bearbeiteter Items. Abbildungstyp 1 – Variante „Gesamt-bearbeitet“ verdeutlicht, wie gut die eigene Klasse (bzw. Gruppe) hinsichtlich der durch-

schnittlichen Anzahl bearbeiteter Items in Relation zu anderen getesteten Klassen (bzw. Gruppen) desselben Leistungsgruppenniveaus ist. Abbildungstyp 2 – Variante „Gesamt-bearbeitet“ gibt darüber hinaus genau aufgeschlüsselt an, welcher relative Anteil von Schülern der eigenen Klasse (bzw. Gruppe) eine bestimmte Anzahl von Items bearbeitet hat (wiederum im Vergleich zum eigenen Leistungsgruppenniveau – bei der 4. Schulstufe im Vergleich zu allen getesteten Klassen in Österreich). Bei dieser zweiten Abbildung wird deutlich, ob sich der Großteil der eigenen Schüler nahe beim Durchschnittswert befindet oder ob die Bearbeitungshäufigkeiten breit streuen. Es kann davon ausgegangen werden, dass sich die unterschiedlichen Anzahlen von ausgewerteten Items in verschiedenen Testformen nur unwesentlich auf die Verteilung in dieser Abbildung auswirken.

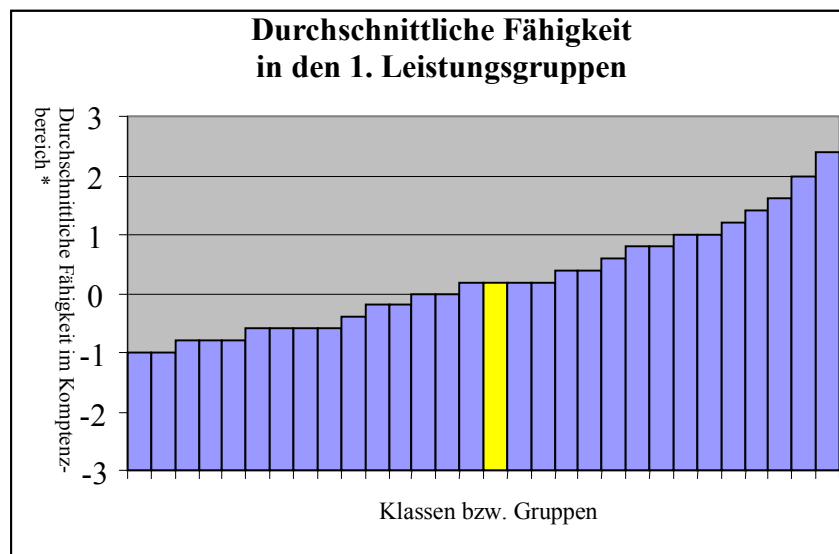
Bezüglich der Anzahl gelöster Items werden zwei analoge Abbildungen dargeboten, die sich von den beiden eben berichteten nur dadurch unterscheiden, dass auf der Abszisse nicht die Anzahl bearbeiteter Items, sondern die Anzahl gelöster Items aufgetragen ist.

Für die einzelnen Kompetenzbereiche werden neben den Bearbeitungs- und Lösungshäufigkeiten dort, wo dies aufgrund der testtheoretischen Analysen möglich ist, die bereits beschriebenen Prozenträge klassen- (bzw. gruppen-) bezogen rückgemeldet. D.h., ein Prozenrang von beispielsweise 40 drückt aus, dass 40 % aller anderen getesteten Klassen (bzw. Gruppen) desselben Leistungsgruppenniveaus durchschnittlich schlechter oder gleich gut in Bezug auf die jeweils erfasste Fähigkeit abgeschnitten haben - diese Fähigkeit ergibt sich aus den Auswertungen nach dem Rasch-Modell, wobei eben Schüler fair vergleichbar sind, selbst wenn sie jeweils andere Items vorgegeben erhielten (vgl. dazu die Ausführungen zum Rasch-Modell im Anhang).

Zusätzlich wird bezüglich der einzelnen Kompetenzbereiche (dort, wo testtheoretische Analysen es erlauben) die durchschnittliche, z.B. mathematische Fähigkeit der Schüler der eigenen Klasse (bzw. Gruppe) mit derjenigen aller anderen Klassen (bzw. Gruppen) desselben Leistungsgruppenniveaus grafisch verglichen (Abbildungstyp 1 – Variante „Fähigkeiten“) – vgl. am Beispiel der 1. Leistungsgruppe, und zwar für M8:

Abbildungstyp 1 – Variante „Durchschnittliche Fähigkeit“: Vergleich der Testleistungen mit allen anderen 1. Leistungsgruppen in Österreich

Sie sehen hier die durchschnittliche Fähigkeit der Schüler/innen in Ihrer 1. Leistungsgruppe im Kompetenzbereich *, gegenübergestellt zu allen anderen 1. Leistungsgruppen in Österreich. Jede Säule repräsentiert eine bestimmte 1. Leistungsgruppe in Österreich. Je höher die Säule, desto höher ist die durchschnittliche Fähigkeit. Ihre 1. Leistungsgruppe ist andersfarblich hervorgehoben.



Aus der Abbildung wird ersichtlich, wie die eigene Klasse (bzw. Gruppe) im Vergleich zu allen anderen Klassen (bzw. Gruppen) desselben Leistungsgruppenniveaus abschneidet.

Erlauben die testtheoretischen Analysen für einen bestimmten Kompetenzbereich keine Berechnung solch entsprechender Fähigkeiten, dann erfolgt ein grafischer Vergleich mit den durchschnittlichen Testleistungen aller anderen Klassen (bzw. Gruppen) desselben Leistungsgruppenniveaus lediglich über die Anzahl gelöster Items.

Schließlich gibt eine weitere Grafik (analog zu Abbildungstyp 2, vgl. oben) genau aufgeschlüsselt an, welcher relative Anteil von Schülern der eigenen Klasse (bzw. Gruppe) eine bestimmte Anzahl von Items gelöst hat (wiederum im Vergleich zum jeweiligen Leistungsgruppenniveau). Diese Abbildung verdeutlicht, ob sich der Großteil der eigenen Schüler nahe beim Durchschnittswert befindet oder ob die Bearbeitungshäufigkeiten breit streuen. Es kann wieder davon ausgegangen werden, dass sich die unterschiedlichen Anzahlen von ausgewerteten Items in verschiedenen Testformen nur unwesentlich auf die Verteilung in dieser Abbildung auswirken.

Die für manche Unterrichtsfächer erhobenen Angaben zum Unterricht der Schüler (z.B. für M8: „Ich bekomme in Mathematik Nachhilfe“ oder „Ich habe Freude an Mathematik“) werden ebenfalls grafisch aufbereitet; hier erfolgt ein Vergleich der eigenen Klasse (bzw. Gruppe) mit dem entsprechenden Leistungsgruppenniveau (im Fall der 8. Schulstufe) sowie mit allen in Österreich getesteten Schülern.

3.3. Ergebnismeldung auf Schulleiterebene

Schulleiter erhalten je Leistungsgruppenniveau bzw. bei der 4. Schulstufe für diese einen eigenen Login-Code. Der Schulleiter bekommt keine Klassentestergebnisse, sondern nur das Gesamtergebnis in der Schule. Dementsprechend werden auch die einzelnen Testergebnisse mit denen anderer Schulen verglichen. Abgesehen davon werden in der Rückmeldung dieselben Angaben wie auf Lehrerebene gemacht (s. Abschnitt 3.2.).

3.4. Ergebnismeldung auf Schulaufsichtsebene

Die Rückmeldung auf Schulaufsichtsebene erfolgt je Bundesland. Die Schulaufsicht erhält je Leistungsgruppenniveau bzw. für die 4. Schulstufe über alle Schulen einen eigenen Login-Code. Sie bekommt keine Schulergebnisse, sondern nur das Gesamtergebnis im jeweiligen Bundesland. Dementsprechend werden auch die einzelnen Testergebnisse mit denen anderer Bundesländer verglichen. Abgesehen davon werden in der Rückmeldung dieselben Angaben wie auf Lehrer- und Schulleiterebene gemacht (s. Abschnitt 3.2.; die Angabe von Prozenträngen entfällt hier allerdings).

4. Grenzen der Interpretierbarkeit

Verschiedene Aspekte müssen bezüglich der Interpretation der Testergebnisse berücksichtigt werden:

- Es handelt sich bei den Standard-Tests um punktuelle Tests. D.h., die Aussagen sind unter dem Gesichtspunkt zu relativieren, dass zum ausgewählten (Zeit-) Punkt bestimmte Rahmenbedingungen untypisch (gewesen) sein könnten. So mag ein betreffender Schüler am gegebenen Tag aus irgendwelchen Gründen leistungsmäßig beeinträchtigt gewesen sein, etwa unausgeschlafen.
- Solche Tests weisen, wie alle Messinstrumente (z.B. auch eine Uhr), gewisse Messfehler auf. Dabei sind diese zumeist deutlich größer als bei physikalischen Messinstrumenten. Man muss z.B. berücksichtigen, dass es nur wenige Items sind, mit denen man eine bestimmte Fähigkeit (einen bestimmten Kompetenzbereich) messen will. So können einige Items darunter sein, die für einen bestimmten Schüler weniger passen, z.B. weil ein Item ein Thema beinhaltet, das ihn ganz wenig, alle anderen Schüler aber sehr interessiert. Denkbar ist auch, dass einem Schüler bei einem Item ein Flüchtigkeitsfehler passiert. Die psychologische Testtheorie hat aber natürlich Möglichkeiten, solche Messfehler mit ein zu kalkulieren. Daher gilt eigentlich nicht das genau beobachtete Testergebnis allein, sondern auch ein ganzes Intervall um das Testergebnis herum. Je nachdem ist dieses Intervall einmal größer, einmal kleiner. Vor allem für die Rückmeldung auf Schülerebene kann nur eine grobe Unterscheidung zwischen den Schülern erfolgen, weil der einzelne Schüler lediglich einen kleinen Ausschnitt der vorhandenen Items bearbeitet hat, und die damit einhergehende Messgenauigkeit niedriger ausfällt als bei einer fachpsychologischen Eignungsuntersuchung, zum Beispiel bei Schul- und Bildungsberatungen der Schulpsychologie. Allerdings stehen beim Standard-Test vor allem Gruppenaussagen im Vordergrund, also beispielsweise Aussagen über einzelne Schultypen, bei denen dann von ausreichend hoher Messgenauigkeit auszugehen ist.

- Der Standard-Test dient dem (Schulsystem-) Monitoring und ist mit keinen unmittelbaren Konsequenzen für die Schüler verbunden. So könnte der eine Schüler mehr als andere motiviert sein, ein möglichst gutes Testergebnis zu erhalten.

Es wird natürlich versucht, solchen Grenzen bestmöglich entgegenzuwirken. So wird vor allem die Durchführung (d.i. Testinstruktion und -administration) weitestgehend standardisiert. Bei der Entwicklung der Items wird auch versucht, diese so zu gestalten, dass sie letztlich wirklich nur diejenige Kompetenz erfassen, welche für das jeweilige Fach relevant sind, aber keine darüber hinausgehenden anderen Fähigkeiten und Fertigkeiten für die Lösung ausschlaggebend sind. Insbesondere mit dem Rasch-Modell ist dies empirisch prüfbar, wobei die Messgenauigkeit umso höher wird, je mehr Items ihm tatsächlich entsprechen. Um die Messgenauigkeit zu optimieren, werden für die 8. Schulstufe zusätzlich die Testformen hinsichtlich ihrer Schwierigkeit dem erwarteten Fähigkeitsniveau der Schüler gemäß Leistungsgruppenniveau angepasst. Schließlich wird versucht, die Items für die Schüler möglichst ansprechend und alltagsrelevant zu gestalten.

5. Verwertung der Testergebnisse¹¹

Die Testergebnisse zeigen keinen direkten Weg zur Erhöhung der geforderten Kompetenzen, sondern bieten Anlass für mögliche Optimierungsprozesse. Im Anschluss an eine differenzierte Analyse und Interpretation der Testergebnisse sollte daher die Reflexion über das eigene Handeln im Unterricht erfolgen, mit dem Ziel, den Lehr-Lernprozess noch zu verbessern. Somit geht es um Fördermaßnahmen. Von der Umsetzung konkreter Maßnahmen im Unterricht wird eine Steigerung der Unterrichtsqualität erwartet, die sich in den Leistungen der Schüler niederschlägt. Um Gewissheit darüber zu erhalten, ob die erhofften Verbesserungen auch eingetroffen sind, müssen dann wiederum die Kompetenzen mittels Standard-Test gemessen und beurteilt werden.

Diesen Prozess in Gang setzen sollen „Rückmeldelehrer“. Eigens darin ausgebildete Lehrer besuchen die jeweils erfassten Schulen und helfen bei der Interpretation der schulbezogenen Testergebnisse in einer gemeinsamen Besprechung mit dem Schulleiter und den Lehrern, deren Schüler an der Testung teilnahmen. Individuelle Klassen- bzw. Gruppen-Testergebnisse stehen dabei nicht zur Diskussion, obwohl jeder Lehrer auch eine persönliche Beratung durch den Rückmeldelehrer im Zweiergespräch erbitten kann. Dabei ist nicht vorgesehen, dass die Rückmeldelehrer Maßnahmen vorschlagen. Diese sollen vielmehr in einem Reflexionsprozess an der Schule nach der Rückmeldeinterpretation angedacht werden, wobei von der Projektleitung mittelfristig diverse Hilfestellungen ausgearbeitet und angeboten werden. Die wesentliche Funktion der Rückmeldelehrer beschränkt sich also darauf, dass die Ergebnissrückmeldung für Lehrer und Schulleiter nicht nur über eine Plattform,

¹¹ Vgl. http://www.klassezukunft.at/statisch/zukunft/de/arbeitsbericht_bildungsstandards_14_02_2006.pdf [26.05.2006]

sondern persönlich erfolgt, um zu gewährleisten, dass sie authentisch interpretiert werden, und um den Reflexionsprozess garantiert einzuleiten.

6. Frequently Asked Questions

(Lucyshyn, 2006)

- **Was bedeutet es, wenn sich herausstellt, dass Österreich seine eigenen Standards nicht erfüllt?**

Derzeit kann dies nicht passieren, da sich die Standards noch in einem Entwicklungsprozess befinden und kein Fixum sind, das vorgegeben ist. Standards wollen Erwartungen an die Zukunft ausdrücken und nicht einen Ist-Stand fortschreiben. Das Ergebnis einer Standardüberprüfung soll einen Reflexionsprozess initiieren, dem Maßnahmen zur Qualitätssicherung folgen.

- **An den Standards haben hauptsächlich Lehrer mitgearbeitet; warum nicht „Abnehmer“, die die Ansprüche kennen?**

Erstens sind diejenigen Lehrer, welche an der Entwicklung der Standards mitgearbeitet haben, Experten auf ihrem Gebiet und sehr offen für die Ansprüche der Wirtschaft bzw. der nachfolgenden Ausbildung der Schüler. Zweitens haben die Pflichtschulen die Aufgabe bzw. das Ziel, den Schülern das Rüstzeug für eine weitere Ausbildung zu geben, und nicht schon Spezialfertigkeiten, die von der Wirtschaft sehr differenziert erwartet werden.

- **Können Lehrer, deren Schüler die Standards nicht erfüllen, belangt werden?**

Nein, da das jeweilige Gruppentestergebnis nur der betroffene Lehrer erhält, Schulleiter und Schulaufsicht dagegen nur über die Schule bzw. über das Bundesland aggregierte Testergebnisse. Die Testergebnisse sollen dem Lehrer Auskunft geben, inwieweit er mit seiner Klasse bzw. Gruppe die angestrebten Kompetenzen erreicht hat und auf welchen fachlichen Teilkompetenzen seine Schüler diese nicht erreichen konnten. Die Testergebnisse sollen die Lehrer in ihrer Einschätzung unterstützen, nicht sie kontrollieren.

- **Wie wird ein „Ranking“ vermieden?**

Seitens des BMBWK werden die Daten nicht veröffentlicht. Da nur die jeweils betroffenen Ebenen (Personen) ihre eigenen Daten erhalten, ist ein Ranking schwerlich möglich.

- **Werden alle Schüler überprüft?**

Derzeit wird ein Prüfungszyklus angestrebt, bei dem jede Schule ca. alle 4 bis 5 Jahre getestet wird. Diese Zeit ist notwendig, damit eine Schule die Testergebnisse gewissenhaft aufarbeitet und entsprechende Maßnahmen zur Verbesserung der Qualität setzt. Vom Testen allein wird die Qualität des Unterrichts nicht besser.

- **Erwartet man sich durch die Standards eine Verbesserung der PISA-Ergebnisse?**

Dies kann zwar eine Folgeerscheinung der Standards sein, ist jedoch nicht vorrangiges Ziel. Langfristig wird die Unterrichtsverbesserung jedoch auch einen Niederschlag im PISA-Ergebnis finden. Außerdem können die Schüler mit der Zeit wahrscheinlich besser mit heute noch ungewohnten Testsituationen umgehen.

- **Welche Veränderungen im Bildungssystem werden angestrebt?**

Bezogen auf die Standards wird ein Wechsel von der derzeitigen INPUT-Steuerung (Schwerpunkt auf Inhalt) zu mehr OUTCOME-Orientierung (was können die Schüler nach bestimmten Bildungsabschnitten tatsächlich) möglich.

- **Gefahr des „Teaching to the Test“**

Die Tests sind so angelegt, dass Kompetenzen überprüft werden und nicht „einstudierbare“ Aufgabenlösungen. Außerdem geht von den Tests keine Bedrohung aus (weder für die Schüler noch für das Lehrpersonal), sodass

sich deswegen der Unterricht nicht auf dieses, alle 4-5 Jahre stattfindende Ereignis hin orientieren wird. Es werden die Testitems nicht veröffentlicht, daher ist die Gefahr des „Teachings to the Test“ relativ gering.

- **Wie schaut die Hilfestellung zur Steigerung der Unterrichtsqualität im Rahmen der Bildungsstandards aus?**

Bundes- und landesweit werden für alle Lehrer Weiterbildungsveranstaltungen angeboten. Über die Möglichkeit temporärer Ressourcenzuweisung zur Unterstützung wird verhandelt.

- **Worin besteht der Anreiz für die Schüler?**

Sie bekommen ein detailliertes Testergebnis zu ihren Kompetenzen rückgemeldet, das ihre Selbsteinschätzung verbessern kann. Ein Bewusstsein über die eigenen Stärken und Schwächen kann geschaffen werden. Die Bereitschaft zur Übernahme von Eigenverantwortlichkeit wird gefördert.

Literatur

- Bundesministerium für Bildung, Wissenschaft und Kunst (2004). *Bildungsstandards für Mathematik am Ende der 8. Schulstufe. Version 3.0*. Wien: BMfBWK.
- Heugl, H. (2004). *Die Startveranstaltungen zu den Bildungsstandards Sek I in den Bundesländern*. Unveröff. Vortragsunterlagen.
- Konrad, K. (1999). *Lernstrategien für Kinder*. Baltmannsweiler: Schneider.
- Kubinger, K. D. (2006). *Psychologische Diagnostik – Theorie und Praxis psychologischen Diagnostizierens*. Göttingen: Hogrefe.
- Lucyshyn, J. (2006). *Moderatoren-Manual*. Unveröff. Manuskript, Salzburg: bm:bwk – Projektmanagement Bildungsstandards.
- Mandl, H. & Friedrich, H. F. (Hrsg.) (2006). *Handbuch Lernstrategien*. Göttingen: Hogrefe.
- Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.). *Leistungsmessungen in Schulen* (S. 17-31). Weinheim: Beltz.

Internetquellen:

- <http://www.gemeinsamlernen.at/index2.asp> [26.05.2006]
- http://www.klassezukunft.at/statisch/zukunft/de/arbeitsbericht_bildungsstandards_14_02_2006.pdf [26.05.2006]
- http://www.klassezukunft.at/statisch/zukunft/de/bildungsstandards_folder.pdf [26.05.2006]
- www.testzentrum.at [26.05.2006]
- www.uni-klu.ac.at/lte [26.05.2006]

Anhang

Rasch-Modell

Kurzfassung

im Wesentlichen ein Ausschnitt aus:

Kubinger, K. D. (2003). Adaptives Testen. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 1-9). Weinheim: Beltz/PVU.

Kubinger, K. D. (2003). Testtheorie, Probabilistische. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 415-423). Weinheim: Beltz/PVU.

Vorbemerkung:

Die Theorie zur Konstruktion psychologischer Tests beinhaltet historisch wie wissenschaftstheoretisch bedingt zwei Ansätze: Die sog. „Klassische Testtheorie“ und die **Probabilistische Testtheorie**. Während erstere wesentlich auf korrelationsstatistischen Betrachtungen aufbaut und ein „deterministisches“, nicht prüfbares Modell darstellt, und zwar in Bezug auf den Zusammenhang von *Testwert* und wahrer Eigenschaftsintensität einer Person, verfolgt letztere diesbezüglich einen wahrscheinlichkeitstheoretischen Ansatz; verbunden ist dieser zumeist mit dem Konzept sog. „spezifisch objektiver“ Vergleiche, d.h. zum Beispiel: das Ergebnis eines Eigenschaftsvergleichs zwischen zwei Personen muss unabhängig von anderen Personen sein und (statistisch) unabhängig davon, welche *Items* eines wohldefinierten Itempools dazu herangezogen werden.

1. Rasch-Modell

Das *Rasch-Modell* beschreibt die Wahrscheinlichkeit, dass Testperson (Tp) v Item i löst („+“), in Abhängigkeit eines *Personenparameters* ξ_v , d.i. die (wahre) Fähigkeit von v , und eines *Itemparameters* σ_i , d.i. die (wahre) Schwierigkeit von i . Das heißt, eine bestimmte Fähigkeit ξ_v bedingt nicht deterministisch, ob es zu einer Lösung kommt oder nicht, sondern nur probabilistisch in der Hinsicht, dass die Lösungswahrscheinlichkeit für größere ξ , bei konstantem σ , ebenfalls größer wird. Konkret besteht das Modell aus folgenden Annahmen:

- Das Ausmaß pro Tp hinsichtlich der (einzigen) interessierenden Fähigkeit sowie der Grad der Schwierigkeit pro Aufgabe ist jeweils durch einen einzigen Parameter zu charakterisieren; d.h., sowohl Fähigkeit als auch Schwierigkeit kann ein(!)-dimensional gemessen werden.
- Die Reaktionen jeder Tp über alle Aufgaben hinweg sind „lokal stochastisch unabhängig“: Ob eine bestimmte Tp eine bestimmte Aufgabe löst oder nicht löst, hängt nur von ihrer Fähigkeit und der Schwierigkeit der Aufgabe ab, nicht aber davon, welche anderen Aufgaben sie bereits gelöst hat oder noch lösen wird.
- Als zugrunde liegende Wahrscheinlichkeitsfunktion wird die „logistische“ postuliert, mit additiv zusammengesetzten Parametern ξ und σ :

$$P(+|\xi_v, \sigma_i) = \frac{e^{\xi_v - \sigma_i}}{1 + e^{\xi_v - \sigma_i}}$$

- die Wahrscheinlichkeit für „-“, also dafür, dass Tp v Item i nicht löst, resultiert als Komplementärwahrscheinlichkeit.

Unmittelbar aus der Formel ist abzuleiten:

- Die Wahrscheinlichkeit für eine Aufgabenlösung strebt gegen 1.0 bei immer größeren Fähigkeiten ξ und/oder bei immer kleineren Schwierigkeiten σ ($-\infty \leq \sigma \leq \infty$).
- Die Wahrscheinlichkeit für eine Aufgabenlösung strebt gegen 0.0 im umgekehrten Fall, das ist bei immer kleineren Fähigkeiten ξ ($-\infty \leq \xi \leq \infty$) und/oder bei immer größeren Schwierigkeiten σ .
- Sind ξ_v und σ_i gleich, d.h. entspricht die Aufgabenschwierigkeit dem individuellen Leistungsniveau, dann beträgt die Lösungswahrscheinlichkeit .50.

! Die Bedeutung des *Rasch-Modells* für die *Psychologische Diagnostik* ist in folgendem begründet: Wenn ein Test die Verrechnung der Testleistungen derart vorschreibt, dass als Testwert lediglich die Anzahl gelöster Aufgaben bestimmt werden muss – der Testwert also unabhängig davon ist, welche Aufgaben von der Testperson gelöst wurden und welche nicht –, dann müssen die Aufgaben des Tests *notwendigerweise* diesem Modell folgen, um zu garantieren, dass dieser Testwert tatsächlich die gesamte relevante Information in Bezug auf die fragliche Fähigkeit der Testperson ausschöpft (ein Beweis dieses Gesetzes findet sich bei Fischer, 1974).

Weil sich das *Rasch-Modell* als im statistischen Sinn *stichprobenunabhängig* herausstellt, kann auch ein besonderer Modelltest abgeleitet werden – somit muss es nie ungeprüft vorausgesetzt werden! In letzter Konsequenz bedeutet die „Stichprobenunabhängigkeit“ folgendes: Der Vergleich je zweier Items, zum Beispiel i und j , bezüglich ihrer Schwierigkeiten, σ_i und σ_j , ist unabhängig davon, welche Personenstichprobe dafür verwendet wird – bei der Schätzung dieser Parameter spielt die Wahl der Stichprobe aus einer bestimmten Population für die statistische Inferenz keine Rolle („Spezifische Objektivität“ der Vergleiche). Diese Eigenheit des Modells zieht nun die Idee nach sich: Würde für einen bestimmten Test bzw.

Datensatz das *Rasch-Modell* gelten, so müssten die Parameterschätzungen $\hat{\sigma}$ in verschiedenen Teilstichproben statistisch gleich sein; stellt sich jedoch empirisch heraus, dass wenigstens für ein Item diese Parameterschätzungen nicht gleich sind, dann folgt – per Umkehrschluss: logisch – dass das *Rasch-Modell* nicht gilt. Irgendeine seiner Annahmen ist dann verletzt. Jedenfalls wäre die Anzahl gelöster Aufgaben kein aussagekräftiger Testwert. Diese Anzahl würde vielmehr relevante Information bezüglich der fraglichen Fähigkeit einer Person ignorieren bzw. wäre sie schlicht sinnlos: Gleiche Testwerte drücken nicht gleiche Fähigkeiten aus.

Beweis:

Für den Spezialfall eines Tests mit nur zwei Items ist die Stichprobenunabhängigkeit des *Rasch-Modells* leicht zu beweisen – wenn auch der Beweisansatz, wie bei mathematischen Beweisen häufig, zunächst „aus der Luft gegriffen“ anmutet. Dazu seien die Betrachtungen auf diejenigen Testpersonen beschränkt, welche exakt ein Item gelöst haben (Anzahl gelöster Aufgaben $S = 1$). Die Wahrscheinlichkeit dafür, dass diese Personen Item i , nicht aber Item j lösen, beträgt dann zum Beispiel für Testperson v :

$$\begin{aligned}
P(+, - | S = 1; \xi_v, \sigma_i, \sigma_j) &= \frac{P(+, - | \xi_v, \sigma_i, \sigma_j)}{P(+, - | \xi_v, \sigma_i, \sigma_j) + P(-, + | \xi_v, \sigma_i, \sigma_j)} = \\
&= \frac{P(+ | \xi_v, \sigma_i) \cdot P(- | \xi_v, \sigma_j)}{P(+ | \xi_v, \sigma_i) \cdot P(- | \xi_v, \sigma_j) + P(- | \xi_v, \sigma_i) \cdot P(+ | \xi_v, \sigma_j)} = \\
&= \frac{1}{1 + \frac{P(- | \xi_v, \sigma_i) \cdot P(+ | \xi_v, \sigma_j)}{P(+ | \xi_v, \sigma_i) \cdot P(- | \xi_v, \sigma_j)}} = \\
&= \frac{1}{1 + \left(1 - \frac{e^{\xi_v - \sigma_i}}{1 + e^{\xi_v - \sigma_i}}\right) \cdot \frac{1 + e^{\xi_v - \sigma_i}}{e^{\xi_v - \sigma_i}} \cdot \frac{e^{\xi_v - \sigma_j}}{1 + e^{\xi_v - \sigma_j}} \cdot \left(1 - \frac{e^{\xi_v - \sigma_j}}{1 + e^{\xi_v - \sigma_j}}\right)^{-1}} = \\
&= \frac{1}{1 + \frac{e^{\xi_v - \sigma_j}}{e^{\xi_v - \sigma_i}}} = \frac{1}{1 + e^{\sigma_i - \sigma_j}}
\end{aligned}$$

D.h., die Wahrscheinlichkeit, dass Testperson v , mit genau einer gelösten Aufgabe, gerade Aufgabe i löst, ist unabhängig von ξ_v , also für alle Personen gleich! Die Höhe dieser Wahrscheinlichkeit hängt allein vom Unterschied der Schwierigkeiten der beiden Items i und j ab: Wenn zum Beispiel σ_i wesentlich kleiner als σ_j ist, wird diese Wahrscheinlichkeit groß; wenn beide Aufgaben gleich schwierig sind, beträgt diese Wahrscheinlichkeit exakt .50. Egal welche Fähigkeiten die getesteten Personen in einer beliebigen Stichprobe hatten, die relative Häufigkeit, mit der gerade Item i von Personen mit $S = 1$ gelöst wurde, vermag die Schwierigkeitsdifferenz $\sigma_i - \sigma_j$ laut obigem Ergebnis zu schätzen.

Beispiel:

Der untenstehend angeführte fiktive Fall illustriert die Stichprobenunabhängigkeit des *Rasch*-Modells. Obwohl einmal eine leistungsstarke, das andere Mal eine leistungsschwache Stichprobe zugrunde liegt, resultiert dieselbe Schätzung für $\sigma_i - \sigma_j$ – die Bezeichnungen „leistungsstark“ und „leistungsschwach“ richten sich nach den relativen Lösungshäufigkeiten.

		leistungsstark					leistungsschwach		
		Item i					Item 1		
		+	-				+	-	
Item j	+	750	50	800	Item j	+	80	40	120
	-	150	50	200		-	120	760	880
		900	100	1000			200	800	1000
150 + 50 Personen mit $S = 1$					120 + 40 Personen mit $S = 1$				

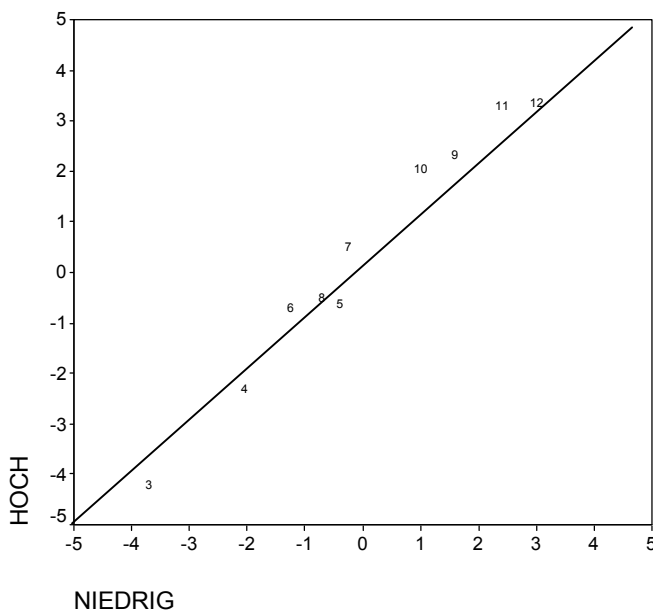
Als Schätzung von $P(+, - | S = 1)$ ergibt sich nämlich aus der leistungsstarken Stichprobe $(150/200)/[(150/200)+(50/200)] = 3/4$ und aus der leistungsschwachen Stichprobe mit $(120/160)/[(120/160)+(40/160)] = 3/4$ dasselbe; somit folgt für die Schätzung von $1/(1 + e^{\sigma_i - \sigma_j})$ beide Male $3/4$, und daraus: $\hat{\sigma}_i - \hat{\sigma}_j = \ln 1/3$.

Ein anschaulicher,

- graphischer Modelltest des *Rasch*-Modells stellt schlicht die Itemparameter, wie sie in zwei verschiedenen (Teil-) Stichproben geschätzt wurden, in einem rechtwinkligen Koordinatensystem gegenüber; gilt das Modell, müssen alle Punkte auf einer 45°-Geraden liegen (die Theorie zur Parameterschätzungen der ξ und vor allem der σ aus empirischen Daten s. z.B. bei Hoijtink & Boomsma, 1995, bzw. Molenaar, 1995). Der bekannteste inferenzstatistische Modelltest ist der sog.
- „(bedingte) *Likelihood-Ratio-Test* von Andersen“. Damit wird geprüft, ob die beobachteten Daten durch die in verschiedenen Teilstichproben separat geschätzten Itemparameter wesentlich besser erklärt werden können als durch die entsprechenden Schätzungen in der Stichprobe insgesamt.

Beispiel:

Die testtheoretischen Analysen des Zusatztests *Unmittelbares Reproduzieren-figural/abstrakt* aus dem AID 2 (Kubinger & Wurst, 2000) ergaben: Unter α -Adjustierung liefern die vier berechneten *Likelihood-Ratio-Tests* für vier Teilkriterien der Gesamtstichprobe (hohe vs. niedrige Anzahl gelöster Aufgaben; männlich vs. weiblich; Deutschland vs. Österreich; bis 10 Jahre vs. über 10 Jahre) nicht signifikante Ergebnisse; die empirischen Daten widersprechen also nicht dem *Rasch*-Modell. Der graphische Modelltest (untenstehend für das Teilkriterium hohe vs. niedrige Anzahl gelöster Aufgaben) verdeutlicht dies: Wie gefordert, liegen die in den beiden Teilstichproben geschätzten und dann gegeneinander aufgetragenen Itemparameter nahezu auf einer 45°-Geraden (vier der 14 Aufgaben haben zu extreme Lösungshäufigkeiten in einer der beiden Teilstichproben, so dass für sie keine zweimalige Parameterschätzung möglich ist).



! Praxis ist es regelmäßig, im Zuge einer Testkonstruktion nach dem *Rasch*-Modell (sukzessive) solche Aufgaben auszuscheiden, die der Modellvoraussetzung der Stichprobenunabhängigkeit widersprechen, und die Modelltests mit dem entsprechend reduzierten Datensatz zu wiederholen. Auf diese Weise gelingt oft eine a-posteriori Modellanpassung. Dann ist es notwendig, die Modellgültigkeit für genau diesen verbleibenden Itempool an Hand einer neu-

en, unabhängigen Stichprobe zu prüfen.

! Im Fall, dass ein Test den Modelltests standhält, ist – auf Grund des „Falsifikationsprinzips“ in der Erkenntnislogik – die Geltung des *Rasch*-Modells natürlich nicht (zwingend) bewiesen. Trotzdem wird sie aber dann üblicherweise als gegeben erachtet. Der „Grad der Bewährung“ *sensu Popper* scheint für das Modell dann ausreichend, wenn tatsächlich mehrere unabhängige Modelltests durchgeführt wurden!

Der faire Vergleich von Testleistungen zwischen zwei oder mehreren Personen, die verschiedene Items bearbeiteten, macht die Anwendung der Probabilistischen Testtheorie notwendig.

! Offensichtlich ist die Anzahl gelöster Items als Testwert ungeeignet: Ein und dieselbe Anzahl, zum Beispiel einmal bei leichten Items, das andere Mal bei schwierigen Items erzielt, würde die faktischen Testleistungen nicht adäquat abbilden. Innerhalb der Probabilistischen Testtheorie ist es jedoch möglich – vorausgesetzt eines der von ihr zur Beschreibung diversen Reaktionsverhaltens angebotenen mathematischen Modelle „gilt“ –, die wahre, aber unbekannte Fähigkeit einer beliebigen Person über die jeweilige Modellgleichung unter Berücksichtigung der getroffenen Itemauswahl zu schätzen.

- Bei gegebenen Itemparametern ist also eine Schätzung von ξ_v auf Grund des modellierten Zusammenhangs von Personen- und Itemparametern möglich.

Beispiel: Angenommen Person v hätte die Items 1, 5 und 9 vorgegeben erhalten und davon Item 1 sowie Item 9 gelöst, nicht aber Item 5; dann beträgt die Wahrscheinlichkeit für genau dieses Testverhalten laut 1-PL Modell

$$P(1^+, 5^-, 9^+ | \xi_v, \sigma_1, \sigma_5, \sigma_9) = \frac{e^{\xi_v - \sigma_1}}{1 + e^{\xi_v - \sigma_1}} \cdot \frac{1}{1 + e^{\xi_v - \sigma_5}} \cdot \frac{e^{\xi_v - \sigma_9}}{1 + e^{\xi_v - \sigma_9}},$$

woraus mit Hilfe der Maximum-*Likelihood*-Methode eine Schätzung $\hat{\xi}_v$ relativ leicht bestimmt werden könnte. Diese wäre unmittelbar mit der analog gewonnenen Schätzung $\hat{\xi}_w$ von Person w zu vergleichen, welche etwa die Items 2 und 3 gelöst haben mag, nicht aber Item 4. Setzt man dieses Beispiel numerisch um, so ergibt sich für $\sigma_1 = -3$, $\sigma_2 = -2$, $\sigma_3 = -1$, $\sigma_4 = 0$, $\sigma_5 = 2$, $\sigma_9 = 3$ mit Hilfe eines einschlägigen Computerprogramms $\hat{\xi}_v = 2.509$, $\hat{\xi}_w = -0.195$; obwohl beide Personen gleich viele Items gelöst haben, erweist sich Person v als deutlich fähiger, weil die von ihr erbrachten Lösungen einer Testung an (durchschnittlich) bedeutend schwierigeren Items entstammen.

Organigramm: Struktur der kooperierenden Teams

