Wolfram Meyerhöfer, Universität Potsdam

Was misst TIMSS? Einige Überlegungen zum Problem der Interpretierbarkeit der erhobenen Daten

Abstract:

Bei der Erstellung und Interpretation mathematischer Leistungstests steht die Frage, was eine Aufgabe mißt. Der Artikel stellt mit der strukturalen oder objektiven Hermeneutik eine Methode vor, mit der die verschiedenen Dimensionen der von einer Aufgabe erfassten Fähigkeiten herausgearbeitet werden können. Dabei werden fachliche Anforderungen, Irritationsmomente und das durch die Aufgabe transportierte Bild vom jeweiligen Fach ebenso erfasst wie Momente, die man eher als Testfähigkeit bezeichnen würde.
Am Beispiel einer TIMSS-Aufgabe wird diskutiert, dass das von den Testerstellern benutzte theoretische Konstrukt kaum geeignet ist, nachhaltig zu beschreiben, was eine Aufgabe misst.
Stichworte: TIMSS, Leistungstests, Testfähigkeit, Hermeneutik

The design and interpretation of aptitude tests in mathematics provoke questions as to what each of the set tasks actually measures. With structural or objective hermeneutics, this article introduces a methodology capable of discerning the various dimensions of skills required for a particular task. Not only does this approach allow for the recognition of the technical requirements of the task, its off-putting factors and the image of the subject conveyed. The methodology is also able to locate the elements addressing the kind of skill that can more accurately be classified as „test ability“. Focusing on an example selected from a TIMSS aptitude test, the discussion seeks to demonstrate that the theoretical construction employed in setting the test is hardly suited to define with any sense of permanence what is measured by each task.
Key Words: TIMSS, Aptitude Tests, Test Ability, Hermeneutics

TIMSS hat in Deutschland eine Debatte über Sinn und Unsinn von großen Vergleichstests im Mathematikunterricht ausgelöst. Eine scharfe Kontroverse darüber, was TIMSS gemessen hat, wurde in der Zeitschrift „Die Deutsche Schule“ geführt. Hier greift Hagemeister (1999) die TIMS-Studie ebenso fundamental an, wie Baumert u.a. (2000) sie entschieden verteidigen.

Hagemeister stellt dabei die den Kern des Testproblems treffende Frage „Was misst diese Aufgabe?“ Ich möchte eine Methode vorstellen, die hilft, umfassendere und tiefere Antworten auf diese Frage zu finden, als Hagemeister sie vorstellt. Dabei werden fachliche Anforderungen, Irritationsmomente und das durch die Aufgabe transportierte Bild vom jeweiligen Fach ebenso erfasst wie Momente, die man eher als Testfähigkeit bezeichnen würde.
Es wird sich zeigen, dass die Baumert-Gruppe eine andere Frage stellt, nämlich „Was **soll** die Aufgabe messen?“ Sie versucht, sie u.a. unter Rückgriff auf von ihr konstruierte „Fähigkeitsniveaus“ zu beantworten. Damit wird untersucht, ob „die eingesetzten Aufgaben tatsächlich eine inhaltlich identifizierbare Dimension naturwissenschaftlicher (oder mathematischer) Kompetenz erfassen“. Die hier vorgestellte Aufgabeninterpretation illustriert, dass das Konstrukt der „Fähigkeitsniveaus“ die durch die Aufgabe gestellten Anforderungen nicht fundiert beschreibt.
Als weiteres Ergebnis deutet sich an, dass dem Problem des in verschiedenen Ländern unterschiedlichen Testcoachings durchaus Beachtung geschenkt werden muss.

INTERPRETATION EINER TIMSS-AUFGABE

Meine Aufgabeninterpretation lehnt sich methodisch an die Objektive Hermeneutik an. Diese Methode wurde von Ulrich Oevermann (z.B. 1979) entwickelt, insbesondere um die latente Bedeutung von Texten freizulegen. Dazu arbeitet man die objektive Textstruktur eines Textes her-

aus. Das ist jener Aspekt von Textbedeutung, der nicht in das Belieben eines subjektiven Textverständnisses gestellt ist: „Nicht „was der Autor sagen wollte“, sondern „was er gesagt hat“, ist Gegenstand der interpretativ-methodischen Erschließung.“ (Wernet 2000)

Die Methode der Objektiven Hermeneutik wurde bisher bei der Beurteilung von schulnahen Testaufgaben noch nicht angewandt: Sie ist eine eher junge Methode. Ausserdem werden Testaufgaben als Forschungsgegenstand erst ernst genommen, seit sie stärkeren Einfluss auf Schulwirklichkeit zu nehmen drohen. Wegen ihrer Leistungsfähigkeit bei der Freilegung latenter Textbedeutung bietet sich die Objektive Hermeneutik hier aber besonders an. Das bei TIMSS und auch bei PISA angewandte Rasch-Modell thematisiert den latenten Charakter der zu testenden Leistungseigenschaft explizit, denn es „basiert auf der Annahme, daß die Wahrscheinlichkeit der Lösung einer Aufgabe von der Ausprägung eines latenten Merkmals bei den untersuchten Personen abhängt (Personenparameter)“ (Bortz 1984, S.154)

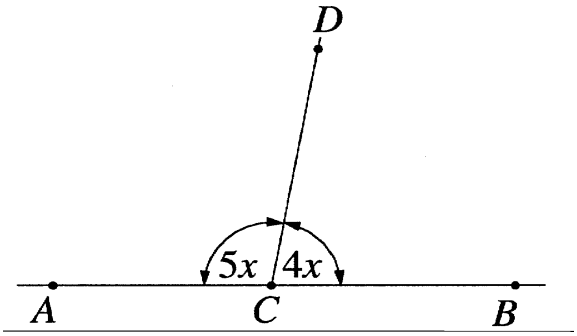
Bei TIMSS wird unterstellt, dass das getestete Merkmal mathematische oder naturwissenschaftliche Leistungsfähigkeit ist. Diese Annahme soll hier hinterfragt werden. Der Vorteil der Objektiven Hermeneutik ist dabei, dass die Interpretation methodisch kontrolliert erfolgt. Hagemeisters Argumentation ist ja u.a. deshalb so wenig überzeugend, weil er ohne jede methodische Strenge arbeitet. Die Baumert-Gruppe setzt dann die vermeintliche Strenge ihrer Parameter dagegen, welche wiederum inhaltlich wenig abgesichert erscheinen.

Ich konzentriere mich hier auf die Streitpunkte der fachdidaktischen Beurteilung und der Frage, inwiefern die Aufgabe Testfähigkeit misst. Hat man die objektive Textstruktur einer Aufgabe erfaßt, dann hat man aber auch starke Anhaltspunkte dafür, wie die Aufgabe konkret auf den Schüler und sein Testverhalten wirkt. Diese Anhaltspunkte geben die Richtung für weitere Untersuchungen mit Schülern an.

Am Beispiel der Aufgabe M7 wird hier eine an der Methode der Objektiven Hermeneutik orientierte Interpretation vorgestellt[1]:

M7. AB ist in dieser Zeichnung eine Gerade.
Wieviel Grad mißt Winkel BCD?

A. 20
B. 40
C. 50
D. 80
E. 100



Aufgabenlösungen

- ACB ist ein gestreckter Winkel. Er ist im Verhältnis 4 zu 5 aufgeteilt. Er ist also aus neun Teilen zusammengesetzt gedacht. Jeder Teil mißt somit 20 Grad. Da der Winkel BCD aus vier solcher Teile besteht, mißt er 80 Grad.

Man könnte auch rechnen: $4x + 5x = 180$, also $9x = 180$, also $x = 20$, also $4x = 80$.

- Eine Möglichkeit der Aufgabenbewältigung ist eine Mischung aus Schätzen und Rückwärtsarbeiten mit Hilfe der multiple-choice-Angebote: Man schätzt, dass der Winkel etwas weniger als 90 Grad misst, und von den angebotenen Lösungen kommt nur 80 Grad in Frage. Andersherum

[1] In diesem Text ist eine vollständige Aufgabeninterpretation aus Platzgründen nicht möglich. Man findet sie unter http://www.math.uni-potsdam.de/prof/o_didaktik/mita/me

kann man gedanklich die Lösungsangebote daraufhin „durchprobieren“, welche am besten zur Zeichnung passt.

- Eine weitere Möglichkeit der Aufgabenbewältigung ist das Messen. Hier ergibt sich ein Winkel von etwa 80 Grad. Diesen Weg kann ich nicht kontrolliert diskutieren, weil Winkelmesser während des Tests nicht zugelassen waren.

Welches Problem meint der Text? Wie ist die Struktur des Problems?

Die Aufgabe stellt - je nach Betrachtungsweise - unterschiedliche Anforderungen.

Wenn man den algebraischen Weg geht, stellt sie vorrangig algebraische, nachrangig geometrische Ansprüche. Äußerlich stellt sie aber eine Geometrieaufgabe dar und wird in der TIMSS-Statistik auch als Geometrieaufgabe geführt. Das Wissen um die Größe von 180 Grad für den gestreckten Winkel ist für den algebraischen Weg der einzige wesentliche geometrische Aspekt der Aufgabe. Die Verflechtung von Geometrie und Algebra stellt für die Aufgabenlösung einen Anspruch und eine Hürde dar. Die modellierende und algebraische Anforderung besteht zunächst im Erkennen der Aufteilung des gestreckten Winkels, sodann im Berechnen der Größe eines Teiles x und (daraus) der Größe des Winkels BCD.

Geht man den Weg der Kombination von Schätzen und Rückwärtsarbeiten, so stehen völlig andere Anforderungen: Man muss Winkelgrößen abschätzen können. Man kann annehmen, dass der Weg hier über den rechten Winkel führt. Man muss also wissen, dass der rechte Winkel 90 Grad misst und dann sehen, dass der gesuchte Winkel etwas kleiner ist. Die Multiple-choice-Angebote müssen dann daraufhin „durchgecheckt“ werden, welches Angebot sich etwas unter 90 Grad bewegt. Hier ist nur ein Angebot vorhanden, nämlich 80 Grad.

Die Aufgabe verweist also auf zwei sehr unterschiedliche Fähigkeitskomplexe, die verschiedenen mathematischen Gebieten entstammen.

Sprachliche Besonderheiten und ihre Bedeutung für die Sinnstruktur der Aufgabe

- „*AB ist*“: Die Bedeutung des „ist“ wird deutlich, wenn man das ebenfalls denkbare „sei“ dagegensetzt. „Sei“ stellt einen Appell an die Vorstellung dar, „ist“ verweist auf etwas, das für den Leser sichtbar bzw. sonstig erfahrbar ist - hier durch eine Zeichnung.

„Sei“ nimmt aber auch einen relevanten mathematischen Gedanken auf: AB „ist“ im Bild gar keine Gerade. AB ist nur gezeichnetes Sinnbild eines Ausschnitts einer Gerade. „Sei“ verweist darauf, daß eine Gerade nur in unserer Vorstellung existiert. Eine Gerade ist unbegrenzt und eindimensional, der Schüler kann sie sich als unendlich lang und unendlich dünn vorstellen. Die Verwendung des „ist“ blendet diese Problematik und damit einen wichtigen mathematischen Gedanken aus.
Nun muss sich der Tester die Frage stellen, ob die Verwendung von „sei“ für Schüler irritierend bzw. „zu schwer“ ist, so dass die mit der Verwendung von „ist“ verbundene Herabsetzung des Schülers als ernst zu nehmender Partner in Sachen Mathematik damit zu begründen ist. Der andere hierbei zu bedenkende Punkt ist die mittlerweile in Schullehrbüchern übliche Ablösung des „sei“ durch das „ist“. An dieser Stelle knüpft die Entscheidung des Testentwicklers für die eine oder andere Version an schulische Gegebenheiten an. Die Verwendung des „ist“ verweist damit nicht auf eine Testspezifität, sondern auf den Umgang mit mathematischen Ideen im schulischen Feld.

- „*in dieser Zeichnung*“: verweist den Leser explizit auf die Zeichnung, führt also zu schnellerem Erfassen des gesamten Problems (Aufgabe und Zeichnung). Der Aufgabentext ist auch ohne diese Sinneinheit denkbar. Lässt man sie gedanklich weg, wird besonders klar, dass es sich um einen Textteil handelt, der zur schnelleren Erfassung der Aufgabenstellung führt.

- *„eine Gerade“*: Mit der Bezeichnung von AB als Gerade ist ein Irritationsmoment gegeben, da es eine Vielzahl von Bezeichnungskonventionen gibt. Dieses Irritationsmoment verbleibt unabhängig davon, für welche Bezeichnungsweise man sich entscheidet, da verschiedene Lehrer mit verschiedenen Bezeichnungskonventionen arbeiten: Einen in geometrischen Bezeichnungsweisen seines jeweiligen Lehrers penibel geschulten Schüler kann eine Bezeichnungsabweichung verunsichern. Im Test bedeutet das Zeitverlust. Das Problem ist allerdings unlösbar. Zur Vermeidung von Irritation hilft hier Testfähigkeit: der testgeschulte Schüler wird sich von solchen Bezeichnungseinzelheiten nicht irritieren lassen.[2]

- Warum wird AB als Gerade eingeführt? Dies beugt einem eventuellen Wahrnehmungsproblem vor: Man könnte ohne diese Angabe aus der Zeichnung heraus die Möglichkeit denken, daß ACB ein anderer als ein gestreckter Winkel ist. Für CD steht ein vergleichbares Problem nicht. Damit wird klar, dass die Einführung von AB als Gerade latent zur algebraischen Lösung führt: Bei einer Kombination von Schätzen und Rückwärtsarbeiten ist dieser Fakt völlig überflüssig.

- Die Winkelbezeichnung und die Bezeichnung der Winkelgrößen mit 4x/5x- verweisen auf den Charakter der Aufgabe als algebraisches Problem. Die Aufgabe wird erst dadurch zu einer Algebraaufgabe, dass die Verhältnisse der Größen der Teilwinkel angegeben werden. Damit müssen 4x und 5x in der Zeichnung eingeschrieben sein: Eine verbale Formulierung wäre schwieriger und würde die Schnelligkeit der Aufgabenerfassung vermindern.

Die Bezeichnung der Winkel mit 4x und 5x birgt zwei Irritationsmomente in sich. Das erste ist die Unüblichkeit der Beschriftung, konventionell ist die Beschriftung mit griechischen Buchstaben. Das zweite ist die Möglichkeit der Lesart „4 mal ... und 5 mal ...“. Von dieser Lesart ausgehend kann man zwar zu einer algebraischen Betrachtung gelangen, sie kann aber auch in eine Sackgasse der Interpretierbarkeit des Problems führen.[3]

- *„Wieviel Grad mißt der Winkel BCD“*: „Mißt“ verweist äußerlich auf Messen - allerdings nicht in jedem Fall, wie die Fragestellung „Wieviel Grad mißt der dritte Winkel eines Dreiecks, wenn die beiden anderen Winkel 60 Grad und 40 Grad sind?“ zeigt: Hier wird man nicht aufgefordert, das Dreieck zu zeichnen und dann zu messen. Erst im Zusammenhang mit dem Charakter als Multiple-choice-Aufgabe mit vielen offensichtlich unsinnigen Antwortangeboten führt der Verweis auf das „Messen“ vom Rechnen weg und auf Schätz-Rückwärtsarbeitestrategien hin. Dies gilt auch dann, wenn man ausschließt, daß die Schüler in den verschiedenen TIMSS-Ländern unter den konkreten Testbedingungen den Winkel einfach ausgemessen haben.

Die alternative Aufforderung „Berechne den Winkel ACD!“ würde explizit auf Berechnen verweisen. Sie hätte die Lösungswegvielfalt eingeschränkt, indem sie Schätz-Rückwärtsstrategien und Verfahren, die Rechnen und Auszählen verbinden, latent behindert.

„Wie groß ist ...“ wäre eine neutrale Aufgabenformulierung gewesen, die alle Lösungsverfahren herausgefordert hätte.

„Bestimme die Größe...“ hätte in stärkerem Maße Vorwärtsarbeiten verlangt und ist dann die angemessene Aufgabenformulierung, wenn auf Vorwärtsarbeiten verwiesen und gleichzeitig die Offenheit der Aufgabenstellung gewährleistet sein soll.

[2] Hier zeigt sich auch die fließende Grenze zwischen Testschulung und anderen, von der derzeitigen Bildungsideologie gewünschten Fähigkeiten: Auch ein in mathematischen Bezeichnungsweisen auf Flexibilität geschulter Schüler wird hier nicht irritiert sein.

[3] Reinhard Woschek (Duisburg) führte eine Untersuchung zur Bearbeitung von TIMSS-Aufgaben durch. Die Schüler sollten ihre Vorgehensweise bei der Aufgabenlösung aufschreiben. Nach seinem Bericht las die Mehrzahl der Schüler die Zeichnung im Sinne von „4 mal ... und 5 mal ...“. In diesem Sinne stellt die Bezeichnung eine Irritationsquelle dar.

Multiple Choice

Die Interpretation des Aufgabentextes zeigt, daß die Textstruktur auf ein schnelles Erfassen der Aufgabenstellung ausgerichtet ist. Der Faktor „Schnelligkeit bei der Aufgabenbewältigung“ findet sich auch in der Anwendung von multiple choice bei der Aufgabenlösung wieder. Welchen Einfluß haben - neben den bereits beschriebenen - die Multiple-choice-Angebote auf den Charakter der Aufgabe bei verschiedenen naheliegenden Bearbeitungswegen?

- Falls der Schüler die Aufgabe über Abschätzen in Verbindung mit Rückwärtsarbeiten löst, ist von den angebotenen Winkeln (20, 40, 50, 80 und 100 Grad) nur jener der Größe von 80 Grad mit der Zeichnung in Einklang zu bringen. Die anderen Winkel sind im Sinne der Zeichnung absurd. Die Multiple-choice-Angebote vereinfachen Abschätzen, weil über mögliche Winkelgrößen und über sinnvolle Genauigkeitsangaben nicht näher nachgedacht werden muss. Sie ermöglichen Abschätzen aber auch erst: Würde man die Angebote 75, 77, 80, 82, 85 Grad machen oder mit einer 7x/3x-Konstruktion arbeiten, so wäre Abschätzen nicht möglich.

- Der Schüler kann das Problem in der eingangs beschriebenen algebraischen Weise lösen. Dazu muss er ACD als gestreckten Winkel mit einer Größe von 180 Grad erkennen. Die Multiple-choice-Angebote helfen ihm nur im Sinne einer Bestätigung. Allerdings sind mit 20 Grad und 100 Grad auch zwei für Deutschland wesentliche Fehlerquellen - der Schüler bestimmt nur x bzw. er betrachtet den falschen Winkel - als Angebote vertreten.[4]

- Der Schüler kann auch rückwärts arbeiten. Das bedeutet, daß er die angebotenen Lösungen viertelt und errechnet, ob 5x den Rest zu 180 Grad ergeben. Dabei ist es sinnvoll, zunächst nur die Angebote zu betrachten, die sich leicht vierteln lassen. Mit „Hinsehen“, also Abschätzen, bietet es sich sogar an, zunächst nur 80 Grad zu betrachten. Man gelangt damit auch sofort zur Lösung. Rückwärtsarbeiten ist also dann effektiv, wenn man die Lösungsmöglichkeiten vorher einschränkt, indem man schätzt oder mit den bequemsten Zahlen zu arbeiten beginnt.

Der Schüler muss im Gegensatz zum Vorwärtsarbeiten nicht erkennen, daß 9x = 180 sind, sondern kann den Gesamtwinkel von 180 Grad schrittweise zusammensetzen. Das Vorhandensein von Multiple-choice-Angeboten ermöglicht also ein Vorgehen, welches völlig andere Anforderungen stellt als Vorwärtsarbeiten oder Schätzen.

Fazit: Die vorwärts arbeitende algebraische Aufgabenlösung läßt sich kurz hinschreiben, ist aber inhaltlich recht schwierig. Der prüfungslogisch und -ökonomisch beste Weg ist der Weg des Rückwärtsarbeitens und Schätzens. Es erhält derjenige eine erhebliche Zeitprämie, der diesen Weg geht. Multiple choice sorgt hier also dafür, dass von vielen denkbaren Lösungswegen ein bestimmter mit Abstand zu größerem Testerfolg führt. Latent und manifest werden aber mehrere Lösungswege nahegelegt.

Fazit zur Interpretation der Aufgabe M7

Die Interpretation läßt sich unter drei Fragestellungen zusammenfassen:

1. Was soll die Aufgabe testen und was testet sie?
2. Wird die Interpretierbarkeit der Testaufgabe durch den Aufgabentext verändert?

---

[4] Was bedeutet es, dass 20 und 100 Grad angegeben sind?

Mit dem Angebot 20 Grad wird der Schüler diagnostiziert, der ohne Berücksichtigung der Frage die deutsche Standardaufgabe „Bestimme x!“ löst. Diese Aufgabe ist rechnerisch weniger umfangreich, aber inhaltlich nicht einfacher.
Der Nebenwinkel von BCD ist 100 Grad groß. Der mathematische Gehalt wird nicht verändert, wenn man die Aufgabe für den „falschen“ der beiden Winkel betrachtet. Hier findet Erziehung zu „Exaktheit“ statt, die allerdings keine mathematische Exaktheit ist, sondern eher Aufmerksamkeit für Nebensächliches.

3. Welcher mathematische Geist wird durch den Aufgabentext transportiert? Dabei verstehe ich unter „Geist von Mathematik“ das vermittelte Bild von Mathematik, die Ernsthaftigkeit der Auseinandersetzung mit dem mathematischen Inhalt und die zur Aufgabenlösung erforderlichen Kenntnisse, Fähigkeiten, Fertigkeiten.

1. Die Aufgabe soll testen, ob der Schüler in der Lage ist,

- entweder ein algebraisches Problem aus einem geometrischen Kontext herauszuschälen, es algebraisch zu bearbeiten und die Lösung geometrisch zu interpretieren
- oder einen Winkel im Vergleich mit mehreren vorgegebenen Maßen abzuschätzen.

Sie testet daneben aber

- Irritationsresistenz
- flexiblen Umgang mit geometrischen Bezeichnungen
- Konzentration auf den gefragten Gegenstand und
- die Fähigkeit, sich gegen einen komplizierten und für einen schnelleren, zwar im Prinzip weniger genauen, die spezifische Anforderung der Testsituation aber abdeckenden Weg entscheiden zu können.

Das Lösungsverhalten lässt bis auf wenige dieser Punkte (20/100-Grad-Antwort) keinen Rückschluss auf die Rolle der einzelnen Parameter für das Lösungsverhalten zu.

Man könnte stark vereinfachend sagen, die Aufgabe testet, ob der Schüler in der Lage ist, die richtige Lösung anzukreuzen. Im englischsprachigen Raum gibt es dafür die Zuspitzung: Tests test tests.

2. Das Testresultat ist durch die gewählte Textkonstruktion kaum interpretierbar.

Der Text fordert latent („mißt“) und durch die Kombination von Zeichnung und vorgeschlagenen Multiple-choice-Alternativen zu einer Kombination von Schätzen und Rückwärtsarbeiten auf.

Der Text fordert aber auch zu einer algebraischen Lösung auf. Dies wird latent durch die Einführung von AB als Gerade, durch gleichzeitiges Weglassen des Verweises auf CD als Gerade und manifest durch die 4x/5x-Konstruktion eingefordert.

Selbst ein reines Schätzen führt - wenn auch unsicher - zur Lösung.

Man kann also über völlig verschiedene Operationen mit völlig verschiedenen Anforderungen zu einer Lösung gelangen. Drei wesentliche Typen von „Lösern“ lassen sich unterscheiden: Der erste verfügt über algebraische Problemlösefähigkeit und rechnet. Der zweite erkennt zwar die algebraische Anforderung und könnte sie erfüllen, erkennt aber, dass es mit einer Kombination von Schätzen und Rückwärtsarbeiten ohne Rechnen schneller geht. Der dritte erkennt die algebraische Anforderung gar nicht oder kann sie nicht bewältigen; aus diesem Unvermögen heraus wendet er eine Schätz-Rückwärtsarbeitestrategie an. Für jeden der drei Typen gibt es spezifische Quellen des Scheiterns.

Aus der richtig angekreuzten Lösung heraus sind die drei Typen nicht zu unterscheiden. Eine Interpretation der angekreuzten Lösung ist also nicht trennscharf: Man kann nicht feststellen, welche Fähigkeit man gemessen hat. Das muß nicht schlimm sein, denn man kann ja auch damit zufrieden sein, irgendeine nicht näher beschriebene „mathematische Fähigkeit“ zu testen. Man muß sich dessen nur bewußt sein.

3. Welcher mathematische Geist wird durch die Aufgabe vermittelt?

Es ist tendenziell ein Geist des Nicht-Ernst-Nehmens. Der mathematische Inhalt wird nicht ernst genommen, und der Schüler wird als Experte nicht ernst genommen. Diese Logik findet sich zunächst in der Verwendung des „ist“ statt des „sei“, hier wird mit dem Gedanken der Verständlichkeit der Abstraktionscharakter der Geraden ignoriert. Sie findet sich auch in der Verschiebung der Prioritäten vom mathematischen Kern zu einer Aufmerksamkeitsprüfung durch das 20- und das 100-Grad-Angebot. Hier spiegelt sich im Test aber zu großen Teilen lediglich der Geist des derzeitigen Mathematikunterrichts.

Die Hinführung auf eine Methode des Schätzens und des Rückwärtsarbeitens vermittelt aber auch einen Geist von Mathematik, der über reines Rechnen hinausweist. Auch das 20-Grad-Angebot weist über eine hiesige Einschränkung von Aufgabenstellungen hinaus.

Die Aufgabe illustriert damit ein Dilemma derzeitiger Testkonstruktion: Man nimmt den Geist des vorherrschenden Mathematikunterrichts auf und versucht doch auch, Neuerungen zu transportieren. Sobald Tests eine größere Bedeutung erlangen besteht die Gefahr, dass der Getestete in diesem Spannungsfeld zerrieben wird.

Die Aufgabe M7 eignet sich kaum zum Testen mathematischer Leistungsfähigkeit. Sie eignet sich aber für einen Test, in dem schnelle, clevere Problemlöser von langsameren, dezidierter und kalkülorientierter arbeitenden Personen separiert werden sollen.

WAS SOLL DIE AUFGABE MESSEN?

Die an der Objektiven Hermeneutik orientierte Interpretation hat nicht nur versucht, die Frage „Was misst die Aufgabe?“ zu beantworten. Sie hat - quasi nebenbei - auch eine Aussage dazu getroffen, was gemessen werden *soll*[5]. Sie geht bei letzterem über eine normale fachdidaktische Analyse nicht hinaus, weil sie wie diese dazu nur den manifesten Aufgabentext analysiert.

Die Debatte zwischen Hagemeister und der Baumert-Gruppe wird nun dadurch verschärft und auch unproduktiv, dass ihre unterschiedlichen Fragestellungen nicht thematisiert werden. Hagemeister stellt die Frage „Was wird hier gemessen?“ Die Baumert-Gruppe widmet sich aber der Frage „Was **soll** die Aufgabe messen?“ Diese Fragestellung ist aus der Perspektive dessen, der die Aufgabe erstellt, zunächst auch verständlich. Die Gruppe beschreibt dazu ein „subjektives Situationsmodell“ für die jeweilige Aufgabe. „Mit der Entwicklung des subjektiven Situationsmodells wird entschieden, um was es überhaupt geht, welches Wissen aktiviert, welcher Lösungsweg gewählt und welche Denkoperationen durchgeführt werden. Das Situationsmodell legt auch fest, auf welchem fachlichen Anspruchsniveau die Aufgabe behandelt wird. ... Eine gute Testaufgabe enthält in sparsamer Form die notwendigen Hinweisinformationen, die bei Probanden, die das durch die Aufgabe indizierte Fähigkeitsniveau erreichen oder übertreffen, zur Bildung eines für die Lösung der Aufgabe adäquaten Situationsmodells führen.“ (Baumert u.a. 2000, S. 196)

Hier wird lediglich thematisiert, was gemessen werden soll, und für die Erstellung der Testaufgabe liegt damit auch eine Hilfe vor. Nach Vorliegen der Aufgabe beantworten sowohl Hagemeisters als auch die objektiv-hermeneutische Interpretation diese Frage ebenfalls. Sie fragen aber darüber hinausgehend, ob das auch der Fall ist. Die Mechanismen der TIMSS-Erstellung konnten diese Frage offenbar nicht vollständig beantworten. Die objektiv-hermeneutische Untersuchung zeigt auch, dass ein Experte, der einen Stapel Aufgaben beurteilen soll, dies quasi nebenbei nicht leisten kann, er kann nur oberflächlich urteilen, was für eine aussagekräftige Auf-

[5] Vereinfacht kann man sagen: Die manifeste Textbedeutung verrät uns, was die Aufgabe messen soll, die latente Textbedeutung und ihre Differenz zur latenten Textbedeutung verrät uns, was sie misst.

gabe eben nicht genügt. Deshalb trifft der Vorwurf von Hagemeister (2000) gegen seine Ko-Gutachter auch kaum, „sich seinerzeit zu hastig und oberflächlich die TIMSS-Items angesehen zu haben." Eine Beurteilung „per Hingucken" legt vorhandene Tücken eben nur teilweise frei, und eine ad-hoc-Durchsicht großer Aufgabenmengen vermindert diese Chance zusätzlich. Bei der Aufgabe M7 hätte man mit so einer „Hinguckmethode" wahrscheinlich nur das „Soll" und die 20/100-Grad-Problematik erfasst. Die tiefergehenden Parameter wären nicht erschlossen worden. Die Problematik der Schnelligkeit der zwei wesentlichen „Soll-Lösungswege" wäre nur bei sehr sensiblem Arbeiten klar geworden.

Hinzu kommt, dass bei TIMSS im Vorhinein für die einzelne Testaufgabe nicht klar war, was sie eigentlich messen soll. Die Aufgaben wurden gar nicht von inhaltlichen Überlegungen ausgehend mit Hilfe eines „subjektiven Situationsmodells" konstruiert, sondern aus einem vorhandenen Aufgabenpool zusammengesucht. Auch die Einordnung in „Fähigkeitsniveaus" erfolgte nicht bei der Testerstellung, sondern hinterher. Deshalb ist das von der Baumert-Gruppe gegen Hagemeister für jede einzelne Aufgabe vorgebrachte Argument, sie „solle" ja etwas anderes messen als er glaube, nicht präzise. Mit dem PISA-Test ist man hier einen Schritt weiter gekommen, denn hier wurde zunächst eine Vorstellung dessen, was man überhaupt testen will, erstellt. Dieses Konstrukt der „mathematischen Literalität" wurde in verschiedene Aspekte aufgeschlüsselt und zu diesen Aspekten Testaufgaben entwickelt. Das ermöglicht die präzise Beantwortung der Frage: Testet die Aufgabe, was sie testen soll? Damit wird vielleicht auch die Ignoranz vermindert, die die Baumert-Gruppe der Bedeutung der Items entgegenbringt. So schreibt sie: „Dennoch gibt es in der deutschen Übersetzung eine Reihe von Testaufgaben, bei denen man sich beim dritten und vierten Durchlesen bessere Lösungen vorstellen könnte - auch wenn die Items in ihren Messeigenschaften dadurch nicht tangiert werden." (Baumert u.a. 2000, S.209) Für die Aufgabe M12 (S.201 ff.) arbeitet sie sogar dezidiert heraus, dass der Schüler nichts vom physikalischen Inhalt verstehen muss, um die Aufgabe zu lösen. Die Aufgabe sei „wenig glücklich, auch wenn die üblichen Itemparameter auf keinerlei Mängel hinweisen."(S.206) Mir scheint das ein Alarmsignal zum Hinterfragen der Aussagekraft der Itemparameter zu sein.

GRENZEN DES KOMPETENZMODELLS DER BAUMERT-GRUPPE

Die TIMSS-Kritik von Hagemeister (1999) stützt sich auf eine fachliche Diskussion der Aufgabeninhalte und eine Art ad-hoc-Interpretation von Schüleräusserungen. Die Baumert-Gruppe legt den Schwerpunkt ihrer Gegenargumentation auf ein Verfahren der „systematischen Beschreibung von Kompetenzstufen": „Das von Beaton und Allen vorgeschlagene Verfahren macht sich eine besondere Eigenschaft des testtheoretischen Modells einer Rasch-Skala zunutze, die es erlaubt, die Fähigkeitskennwerte der Bearbeiter und die Schwierigkeits-Kennwerte der Testaufgaben auf derselben Skala anzuordnen. Die Wahrscheinlichkeit, ein bestimmtes Item korrekt zu lösen, steigt mit der Fähigkeit des Bearbeiters an ... Den Schwierigkeitsgrad einer Aufgabe kennzeichnet man nun durch jenen Punkt auf der Fähigkeitsskala, bei dem sie mit einer Wahrscheinlichkeit von 65 Prozent richtig beantwortet wird. Je schwieriger eine Aufgabe ist, desto höher liegt dieser Referenzpunkt auf der Skala." (Baumert u.a. 2000, S.109)

Für die oben interpretierte Aufgabe M7, der ein Rasch-Wert von 457,1 zugeordnet wurde bedeutet das, dass 65% der internationalen Testgruppe, für die der „Fähigkeits"-Wert 457,1 oder schlechter errechnet wurde, diese Aufgabe gelöst haben.

Baumert u.a. (1997, S.78 ff.; 2000, S.109 ff) schauen sich jetzt die Aufgaben mit bestimmten Raschwerten an und konstruieren daraus „Kompetenzstufen", die zur Grundlage ihrer gesamten *inhaltlichen* Interpretation der Testresultate werden. Sie beschreiben die Fähigkeitsniveaus um

die Aufgabe M7 herum z.B. so: „Bei einem Fähigkeitsniveau von 400 („Elementare Rechenkenntnisse“) kann man von einer einigermaßen sicheren Beherrschung der Grundrechenarten ausgehen, solange die Aufgaben ein Operieren ausschließlich mit natürlichen und nicht zu großen Zahlen verlangen. Einfache Tabellen und quantitative Verhältnisse darstellende Piktogramme werden verstanden. ... Bei einem Fähigkeitsniveau von 450 („Intermediäres Niveau I“) erweitert sich das mathematische Repertoire nur wenig. Ein elementares Verständnis von gewöhnlichen Brüchen und Dezimalbrüchen wird erreicht, so daß Größenunterschiede erkannt werden. Gleichnamige Brüche können addiert und subtrahiert werden, solange keine negativen Zahlen auftreten. Einfache algebraische Terme werden verstanden. ... Ein Grundniveau von 500 („Beherrschung von Routineverfahren“) eröffnet das Verständnis für zentrale, vom Lehrplan für die 6. bis 8. Jahrgangsstufe vorgeschriebene Stoffe. Einfache lineare Gleichungen mit einer Unbekannten werden verstanden; Termumformungen gelingen bei einfachen Aufgabenstellungen. ...“ (Baumert u.a. 1997, S.80)[6]

Bei dieser Sichtweise auf die Aufgabe M7 interpretiert man sie mit einem Rasch-Wert von 457 als eine Aufgabe, die mit elementaren Rechenkenntnissen lösbar ist. Die internationale Lösungswahrscheinlichkeit liegt bei 0,67/0,72, die deutsche aber nur bei 0,58/0,68 (jeweils 7./8.Klasse). Der in der 7.Klasse beachtliche Unterschied lässt sich in dieser Sichtweise aber überhaupt nicht interpretieren. Einerseits gibt es den „internationalen Schüler“ nicht, der die Zahlen erzeugt, die hier zu einer inhaltlichen Interpretation herangezogen werden. Andererseits lässt die hier gewählte Multiple-choice-Konstruktion (ausser teilweise interpretierbaren Fehlerhäufigkeiten) keine Aussagen über die Deutung des Ergebnisses zu. Die prinzipielle Schwäche der statistischen Konstruktion mit einer Lösungswahrscheinlichkeit von 65% verwischt jegliche Interpretierbarkeit zusätzlich.

Die Interpretation in Anlehnung an die Objektive Hermeneutik hat uns hingegen nicht nur die Doppelgesichtigkeit der Aufgabe als einfache Schätz-Rückwärtsarbeite- oder als Algebraaufgabe mit erschwerendem geometrischem Aspekt gezeigt. Sie hat auch aufgezeigt, wo genau im Text sich diese Doppelgesichtigkeit festmacht und wo potentielle Irritationspunkte für den Schüler sind. Sie hat weitere Parameter festgemacht, die Einfluss auf die Lösungswahrscheinlichkeit haben. Zu guter Letzt lehrte sie uns auch noch etwas über den Geist von Mathematik, der mit der Aufgabe transportiert wird. All diese Aspekte werden natürlich umso relevanter, je bedeutsamer Testaufgaben für den Unterricht werden. Wenn Testaufgaben für die Beurteilung von Schulen oder Schulsystemen herangezogen werden, muss einerseits völlig klar sein, was sie messen, andererseits sollten die durch die Aufgaben gesetzten Maßstäbe den Testerstellern in all ihren Dimensionen deutlich sein.

Die Frage, wo genau die Ursachen für ein bestimmtes Ergebnis deutscher Schüler liegt, wird natürlich präziser beantwortet, wenn man nun auch noch Schüler befragt. Reinhard Woschek ließ deutsche und Schweizer Schüler ihre Lösungswege während eines Tests mit TIMSS-Aufgaben aufschreiben. Deutsche Schüler gingen die Aufgabe M7 erheblich häufiger algebraisch an, während Schweizer Schüler eher schätzten. Hier deutet sich an, dass in Deutschland der schwierigere und zeitraubendere Lösungsweg häufiger gegangen wird und mehr Schüler an ihm

[6] Fragen der Sinnhaftigkeit, vielleicht Beliebigkeit solcher Zahlenangaben seien hier nur am Rande angeschnitten:

- Wenn wir hexadezimal dächten, dann hätte die Baumert-Gruppe doch sicherlich bei 360 oder 420 Punkten ihre Skalierung begonnen. Wäre die Interpretation eine andere gewesen? Hätte die „Beherrschung von Routineverfahren“ dann bei 480 oder bei 540 Punkten „begonnen“?
- Wären die besonders erfolgreichen asiatischen Länder bei TIMSS II nicht dabei gewesen, dann hätte sich die gesamte Skalierung verschoben, und zwar mit den gleichen Testaufgaben. Hätte dann die „Beherrschung von Routineverfahren“ erst bei 530 Punkten begonnen?
- Wie hätten sich die Interpretationen verändert, wenn man umgekehrt vorgegangen wäre: Man interpretiert Aufgaben inhaltlich und schaut hinterher, welche Fähigkeitswerte sie haben. Dann versucht man, aus den Zahlen und den Inhalten Fähigkeitsniveaus zu konstruieren.

scheitern. Eine besondere Rolle spielt in Deutschland die Bestimmung des x mit 20 Grad, was dann als Lösung angegeben wird. Die Klassifkation der Baumert-Gruppe könnte sich hier bewähren, wenn sie für Deutschland und die Schweiz verschiedene Fähigkeitsniveaus derselben Aufgabe feststellt. Das vermindert allerdings die mangelnde Präzision der 65%-Sichtweise nicht, so dass nur eine Kombination hermeneutischer und statistischer Methoden uns wirkliche Antworten darauf verspricht, was mit den Aufgaben getestet wird. Durch den erforderlichen Aufwand wird klar, dass man nicht jedes Jahr möglichst für jedes Bundesland einen neuen und auch noch aussagekräftigen Test entwerfen kann.

ZUM PROBLEM DES TESTCOACHINGS

Hagemeister äussert die Vermutung, dass Länder mit einer ausgeprägten Testorientierung einen Vorteil bei Untersuchungen wie TIMSS haben. Die Baumert-Gruppe (2000, S.108) hält mit Untersuchungen dagegen, die einen Coaching-Effekt verneinen. Die objektiv-hermeneutisch orientierte Aufgabeninterpretation gibt Hinweise, an welchem Punkt Testcoaching eine Rolle spielen könnte. Bei Aufgabe M7 heisst das zuallererst: Lass dich nicht irritieren! Derjenige, der nicht beachtet, dass die Gerade anders bezeichnet wird als in seinem Unterricht, der die Bezeichnung mit 4x/5x einfach ignoriert oder - wenn er algebraisch arbeitet - nicht über eine so ungewöhnliche Bezeichnungsweise stolpert, wer nicht über Geraden oder über Sinnhaftigkeit von Aufgaben nachsinnt, der wird schnell durch die Aufgabe durchkommen. Wer ein „Gespür“ dafür entwickelt hat, dass hier nicht gerechnet werden muss, der erhält eine Zeitprämie. Dies ist eher eine Test- als eine mathematische Fähigkeit.

Ich habe jene sechs Mathematikaufgaben von TIMSS, die alle Schüler lösen mussten, auf „Testfähigkeit“ untersucht. Bis auf eine Ausnahme enthielten sie alle einzelne Elemente dieser begrifflich bisher noch nicht vollständig erschlossenen „Fähigkeit“. Das scheint mir zumindest darauf hinzudeuten, dass es ein Phänomen wie Testcoaching gibt. Eine tiefgehende Untersuchung von Items und von Schülerarbeitsweisen ist an dieser Stelle angebracht, zumal sich andeutet, dass „Testfähigkeit“ sich mit Fähigkeiten, die in der momentanen Bildungsideologie erwünscht sind (z.B. Flexibilität des Denkens, Schnelligkeit des Denkens, Konzentration auf vermeintlich Wesentliches), überlagert.

FAZIT

Ich habe mit einer an die Objektive Hermeneutik angelehnten Aufgabeninterpretation eine Methode vorgestellt, die sich als leistungsfähig für die Untersuchung und Erstellung von Testaufgaben erweist. Durch die Arbeit mit dieser Methode deutet sich an, dass das von der Baumert-Gruppe vorgelegte Modell der „systematischen Beschreibung von Kompetenzstufen“, welches zur Interpretation der TIMSS-Ergebnisse genutzt wurde, die gemessenen Fähigkeiten nicht differenziert genug und durch die Orientierung an einem imaginären „internationalen Schüler“ eventuell auch inhaltlich nicht korrekt beschreibt. Für die Beurteilung des Testcoaching-Problems ergibt sich mit der Methode die Möglichkeit, Dimensionen einer „Testfähigkeit“ zu erarbeiten.

Literatur

Baumert, Jürgen; Klieme, Eckhard; Lehrke, Manfred; Savelsbergh, Elwin: Konzeption und Aussagekraft der TIMSS-Leistungstests. in: Die Deutsche Schule, 92.Jg.2000, S.102-115 (Teil I), S.196-217 (Teil II)

Baumert, Jürgen; Lehmann, Rainer u.a.: TIMSS - Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde. Leske und Budrich, Opladen 1997

Bortz, Jürgen: Lehrbuch der empirischen Forschung. Springer, Berlin 1984

Hagemeister, Volker: Was wurde bei TIMSS erhoben? in: Die Deutsche Schule, 91.Jg.1999, S.160-177

Hagemeister, Volker: Irrwege und Wege zur Testkultur. in: Die Deutsche Schule, 92.Jg.2000, Heft 1

Oevermann, Ulrich; Tilmann Allert, Elisabeth Konan; Jürgen Krambeck: Die Methodologie einer „objektiven Hermeneutik“ und ihre allgemeine forschungslogische Bedeutung in den Sozialwissenschaften. in: Hans-Georg Soeffner (Hg.): Interpretative Verfahren in den Sozial- und Textwissenschaften. Metzler, Stuttgart 1979, S.352-434

Wernet, Andreas: Einführung in die Interpretationstechnik der Objektiven Hermeneutik. Leske und Budrich, Opladen 2000