© Empirische Pädagogik
Zeitschrift zu Theorie und Praxis
erziehungswissenschaftlicher Forschung
2001, 15 (4), 553-566

Originalarbeiten

Reinhold S. Jäger und Lars Balzer

MARKUS und TIMSS – ein Vergleich

Mathematische Literalität (Mathematics Literacy) wird heute als eine Basisqualifikation der kulturellen Alphabetisierung angesehen (s. Klieme et al., 1997, S. 85f). Es liegt daher nahe, in ähnlicher Weise, wie dies im Rahmen von TIMSS begonnen wurde, auf der Ebene eines einzigen Bundeslandes vorzugehen und eine Einschätzung dieser Basisqualifikation vorzunehmen. Diese Art der Alphabetisierung hat aber zwei Seiten: eine Orientierung am Curriculum eines Bundeslandes (hier Rheinland-Pfalz) und eine cross-curriculare Komponente, wie sie durch TIMSS (vgl. Baumert et al., 1997) umgesetzt wurde. Diese beiden Seiten sollen miteinander verglichen werden. Hierzu werden die entsprechenden Daten aus der MARKUS-Untersuchung herangezogen.

Ähnlich wie bei der QuaSUM (Lehmann et al., 2000) wurde bei der Untersuchung MARKUS u. a. ein Ausschnitt von TIMSS-Aufgaben verwendet, um mehreren Fragestellungen nachgehen zu können. Auf Grund der beschränkten Testzeit war es notwendig, diesen Anteil an Aufgaben auf ein Minimum zu begrenzen. Dieses Minimum war durch eine Menge von 15 Aufgaben gegeben. Mit Hilfe dieser Aufgabenmenge sollten zwei Funktionen erfüllt werden: (a) Zum einen sollte gewährleistet sein, dass das Spektrum der Aufgabenschwierigkeiten – wie es in TIMSS empirisch vorgegeben ist (s. hierzu Baumert et al., 1997) – möglichst breit wiedergegeben wird, so dass eine Verteilung resultiert, welche die Fähigkeiten ebenso von guten wie von eher leistungsmäßig schlechteren Schülern abbilden kann. (b) Zum anderen sollte die Möglichkeit gegeben sein, eine gegenseitige Validierung der beiden Testverfahren MARKUS-C (dieses entspricht dem curriculumvaliden Testteil) und MARKUS-T (dieses entspricht den übernommenen TIMSS-Aufgaben) vorzunehmen.

Es muss davon ausgegangen werden, dass beide Testverfahren nicht grundsätzlich Gleiches erfassen, weil ihr Konstruktionsprinzip anderes geartet ist. Die TIMSS-Aufgaben beruhen zunächst auf einer anderen Aufteilung der Sachgebiete und betonen die „Kontextualisierung mathematischer Operationen“ (Lehmann et al., 2000, S. 39). Sie orientieren sich zugleich an einer Aufgabentaxonomie, welche sich vom Pol „Kenntnisse“ bis hin zur „Lösung anwendungsbezogener und mathematischer Probleme“ erstreckt. Es muss allerdings in Frage gestellt werden, ob diese Taxonomie für eine regelgeleitete Konstruktion von Aufgaben (vgl. Hornke & Harbon, 1984) geeignet ist und für Benchmarkingprozesse (vgl. Jäger 2001) im curricularen Kontext herangezogen werden kann.

Die Aufgaben zu MARKUS-C (vgl. Balzer & Jäger, 2001) folgen dem Rationale des Fundamentum und Additum, um hierbei sowohl dem Curriculum auf der Ebene der Gemeinsamkeiten über alle Bildungsgänge hinweg zu entsprechen, aber auch zugleich den Besonderheiten des jeweiligen Bildungsgangs im Additum gerecht zu werden.

Fragestellungen

Im Rahmen dieses Beitrags wird folgenden Fragestellungen nachgegangen werden:

a) Wie hätten die rheinland-pfälzischen Schülerinnen und Schüler der 8. Klasse abgeschnitten, wenn sie an der ursprünglichen TIMSS in ausreichender Anzahl teilgenommen hätten?

b) Kann MARKUS-C an MARKUS-T validiert werden? Welche Ergebnisse ergeben sich hierbei?

Methode

Hinsichtlich der oben ausgeführten zwei Fragestellungen ergeben sich zwangsläufig verschiedene methodische Zugänge, welche – orientiert an den Teilfragestellungen – jeweils skizziert werden sollen.

Zu Fragestellung a): Wie bereits angedeutet konnte bei MARKUS nur auf eine eingeschränkte Anzahl von Aufgaben aus der ursprünglichen Untersuchung von TIMSS zurückgegriffen werden. Die hierbei ausgesuchten Aufgaben waren zwar unter dem Blickwinkel einer breiten Streuung der Itemschwierigkeiten und damit der repräsentierten Kompetenzstufen einerseits und der Variabilität der behan-

delten Themen andererseits ausgewählt worden, doch basieren alle Berechnungen und Schätzungen innerhalb von MARKUS auf dieser reduzierten Itemanzahl.

Zum anderen musste bedacht werden, dass die statistischen Angaben zu den TIMSS-Aufgaben unter der Prämisse einer Lösungswahrscheinlichkeit von 65% bestimmt wurden, wohingegen die entsprechenden Item- und Personenparameter bei MARKUS-C unter der Bedingung einer Lösungswahrscheinlichkeit von 50% geschätzt wurden.

Um nun zu adäquaten Fähigkeitsschätzungen zu gelangen, wurde in drei Stufen vorgegangen:

a) Als Ausgangspunkt wurden die ursprünglichen Schätzwerte für die Schwierigkeiten (S_{orig}) der 15 ausgewählten Original-TIMSS-Items verwendet.

b) Von diesen Schwierigkeiten wurden die Fähigkeitswerte (F‘) der rheinland-pfälzischen Schülerinnen und Schüler unter Zugrundelegung eines entsprechenden Algorithmus geschätzt.

c) Durch Berücksichtigung der Lineartransformation F‘ = 500 + 100z wurde eine Umwandlung in die ursprüngliche TIMSS-Metrik realisiert. Dies erfolgt unter Zugrundelegung der Formel 1:

Formel 1:

$$\mathbf{F' = 500 + 100 * ((S_{orig.} - 0.214901) / 1.106172)}$$

Zu Fragestellung b): Bei dieser Frage wird eine Korrelationshypothese überprüft. Hierzu wird die Produkt-Moment-Korrelation berechnet. Wegen der zu erwartenden Moderatorwirkung der betreffenden Bildungsgänge werden die differenziellen Validitäten auf der Grundlage des jeweiligen Bildungsgangs bestimmt.

Über den Korrelationsansatz hinaus wird der Frage nachgegangen, ob mit Hilfe von MARKUS eine Vorhersage der TIMSS-Werte möglich ist. Hierzu wird auf das Konstruktionsprinzip von MARKUS zurückgegriffen: Die Vorhersage wird aus der Kenntnis der individuellen Werte der Teile $MARKUS_{Fundamentum}$ und $MARKUS_{Additum}$ geleistet. Hierbei interessiert, ob es Schülerinnen und Schüler gibt, bei denen Über- bzw. Unterschätzungen gemessen an den Residualwerten erzielt werden, und wie diese Teilgruppen zu beschreiben sind. In diesem Zusammenhang werden die Residualwerte klassifiziert, um damit Schülerinnen und Schüler zu identifizieren, die in höherem Maße als andere in dieser Prognose überschätzt werden.

Stichprobe

Grundlage der folgenden Ausführungen ist die im Rahmen der MARKUS-Studie erfolgte Gesamterhebung aller 8. Klassen des Landes Rheinland-Pfalz. Details dieser Population sind bereits im Editorial beschrieben worden. Hierbei wurden Daten aus allen Testteilen und Befragungen verwendet. Hauptorientierungspunkt der Beschreibungsbasis ist die Ebene individueller Daten, d. h. die von einzelnen Schülerinnen und Schülern ohne Berücksichtigung der Klassenzugehörigkeit.

TIMSS: eine Hochrechnung auf der Grundlage von MARKUS-T

Alle Schülerinnen und Schüler haben den TIMSS-Teil von MARKUS (= MARKUS-T) bearbeitet. Das sind insgesamt 15 Items aus der Original TIMS-Studie, welche über die gesamte Schwierigkeitsskala von TIMSS variieren. Der Range der Schwierigkeiten erstreckt sich dabei von 693 mit der Kompetenzstufe 5 bis zu 376 mit der Kompetenzstufe 1.

Durch das Einbeziehen einer Grundmenge von Aufgaben aus TIMSS, in diesem Kontext als MARKUS-T dargestellt, kann die Frage beantwortet werden, welches Niveau die rheinland-pfälzischen Schülerinnen und Schüler erreicht hätten, wenn sie an der ursprünglichen Untersuchung von TIMSS teilgenommen hätten.

Die in MARKUS neu geschätzten Aufgabenschwierigkeiten der aus TIMSS entlehnten Aufgaben stimmen zwangsläufig nicht exakt mit den Aufgabenschwierigkeiten überein, wie sie in TIMSS berechnet worden sind. Vergleicht man nur rheinland-pfälzische Schülerinnen und Schüler auf der Basis von MARKUS-T, so resultieren hieraus keine Probleme. Will man aber Vergleiche mit den Original-TIMSS-Daten herstellen (z. B. für die Beantwortung der oben genannten Frage), so muss man auch die Aufgabenschwierigkeiten aus der TIMS-Studie zu Grunde legen.

Die Festlegung einer 50-prozentigen Lösungswahrscheinlichkeit zur Schätzung von Personenfähigkeiten, wie sie aus der Skalierung zunächst resultiert und wie sie in MARKUS und QUASUM zu Grunde gelegt worden ist, ist willkürlich. TIMSS wählte einen strengeren Maßstab und legte eine 65-prozentige Lösungswahrscheinlichkeit zu Grunde, bei der eine Aufgabe einem bestimmten Fähigkeitsniveau zugeordnet wird. Formal betrachtet ist dies nur eine Transformation auf der Fähigkeitsskala, die Relationen von Personen innerhalb einer Stichprobe bleiben jedoch unverändert. Dennoch muss für adäquate Vergleiche diese Transformation berücksichtigt werden (s. Formel 1).

Die Daten, die mit Hilfe der bereits zuvor im Rahmen des Methodenteiles beschriebenen Vorgehensweisen berechnet wurden (siehe Abbildung 1), deuten darauf hin, dass die Schätzungen der TIMSS-Werte für die Gruppe der bei MARKUS Untersuchten höher ausfallen als bei der repräsentativen Erhebung der deutschen Schülerinnen und Schüler in der TIMS-Studie (Baumert et al., 1997). Dieses Ergebnis gilt gleichermaßen für alle Vergleichsgruppen. Alle Unterschiede (t-Test aus dem Vergleich zweier unabhängiger Stichproben) zwischen den beiden Vergleichsgruppen MARKUS und TIMSS sind statistisch signifikant ($p < .001$).

Bei der Interpretation dieser Ergebnisse sind allerdings verschiedene Aspekte zu beachten, die aus Gründen der wissenschaftlichen Redlichkeit angeführt werden müssen (vgl. Helmke et al., 2001):

1. Die deutschen und internationalen Ergebnisse sind seit 1997 unter anderem durch die Buchpublikationen von Baumert et al. (1997) sowie Baumert, Bos & Lehmann (2000) einer breiten Öffentlichkeit bekannt geworden. Hierdurch wurden auch die Aufgaben hinreichend transparent und auch zum Teil Gegenstand mathematik-didaktischer Überlegungen und Auseinandersetzungen. Dieser Effekt des Bekanntheitsgrades der Aufgaben muss bei einem solchen Vergleich in Rechnung gestellt werden.

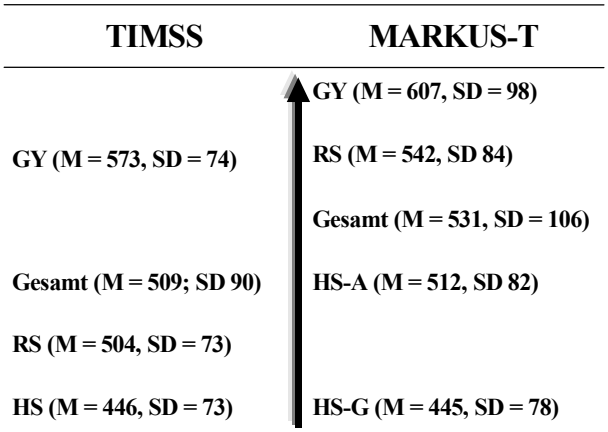
2. MARKUS-T war auf 15 Aufgaben beschränkt, der allerdings die gesamte Breite der Fähigkeitsniveaus abdeckt. Gegenüber TIMSS wurde aber nur ein Ausschnitt der ursprünglichen Aufgaben zur Bearbeitung vorgegeben, so dass der Belastungsgrad der Schülerinnen und Schüler bei MARKUS-T geringer als bei TIMSS war, auch wenn dem MARKUS-T-Teil der Untersuchung eine 45-minütige Testung durch MARKUS-C voraus ging. Die beiden Untersuchungen und deren Bedingungen sind insgesamt nur bedingt vergleichbar. Die Erhebungszeitpunkte liegen fünf Jahre auseinander, die Testbedingungen wie Testdauer, Lehrer als Testleiter, Bearbeitungszeit pro Item, Gesamttestlänge, Speedtestkomponenten u. a. waren substanziell verschieden und erschweren deshalb einen exakten Vergleich.

3. Da bei dieser Untersuchung nur 15 Items aus TIMSS herangezogen wurden, ist eine geringere empirische Basis für die Schätzung von Personenparametern gegeben.

4. Der Vergleich, der nunmehr angestellt wurde, ist ein ausschließlich rechnerischer. Auch wenn empirische Daten zu Grunde liegen, so darf doch nicht übersehen werden, dass zwischen dem ursprünglichen TIMSS-Test und MARKUS-T die genannten Unterschiede bestehen (s. o.). Deshalb gilt der Vergleich nur unter der Bedingung, dass alle angeführten Beschränkungen

keinen bedeutsamen Einfluss hatten. Der Vergleich ist deshalb als eine Art Simulationsstudie anzusehen, nämlich „was wäre gewesen, wenn...“. Den Grad des Einflusses all dieser Bedingungen auf das Testergebnis zu bestimmen, ist allerdings nicht möglich. Es kann daher aber insgesamt vermutet werden, dass die wahren Werte von TIMSS eher etwas überschätzt werden. Aber selbst unter dieser Bedingung sind die rheinland-pfälzischen Schülerinnen und Schüler mit ihren Leistungen nicht schlechter als diejenigen Schülerinnen und Schüler, welche ursprünglich aus der Bundesrepublik Deutschland teilgenommen haben.

Abbildung 1: Vergleich TIMSS – MARKUS-T



Die Validierung von MARKUS-C durch MARKUS-T

MARKUS-C besteht (s. Balzer & Jäger, 2001) aus den Teilen des Fundamentum und Additum. Ein Teil von MARKUS-T wurde dazu verwendet, Grundlagen für einen Vorwissenstest (MARKUS-V) zu bilden. Diese Auswahl kam durch Personen zu Stande, welche als Experten die partizipative Entwicklung des MARKUS-Mathematiktests begleitet haben. Die betreffenden Items gelten als Indikatoren für das Vorwissen und sind somit als Grundbestandteil und Voraussetzung für das Curriculum der 8. Klasse anzusehen.

Unter diesen Bedingungen liegt es nahe, alle genannten Tests korrelationsstatistisch in Beziehung zu setzen. Die nachfolgende Tabelle 1 beschreibt den Zu-

sammenhang zwischen den Tests auf der Grundlage von Produkt-Moment-Korrelationen. Schon allein auf der Basis der Stichprobengröße sind alle angegebenen Korrelationen statistisch signifikant ($p < .001$ für H_0: $r = 0$).

Interessant sind daher ausschließlich die Effektgrößen, bestimmt durch r^2. Gemessen an den Vorgaben von Bortz & Döring (1995, S. 568) gelten die angegebenen Werte als mittlere bis große Effektgrößen. Dieses Faktum gilt unbeschadet der jeweiligen AV bzw. UV. Interessant ist hierbei, dass die geringe Anzahl von Vorwissensitems ausreicht, um gegenüber MARKUS-C immerhin etwas mehr als 30% gemeinsamer Varianz zu erklären.

Tabelle 1: Zusammenhänge zwischen Teiltests von MARKUS

Tests	**MARKUS-V**	**MARKUS-T**
MARKUS-C	.55	.61
MARKUS-Fundamentum	.47	.50
MARKUS-Additum	.48	.54

Interessant ist es dabei der Frage nachzugehen, ob der in Tabelle 1 vorgefundene Zusammenhang zwischen MARKUS-V und MARKUS-C durch den jeweiligen Bildungsgang (s. Balzer & Jäger, 2001) moderiert wird. Die entsprechenden Daten finden sich in Tabelle 2. Hieraus ist zu entnehmen, dass zunächst durch den Bildungsgang kein inkrementeller Zuwachs der Validität zu verzeichnen ist und dass darüber hinaus die Zusammenhänge in allen Bildungsgängen eine vergleichbare substanzielle Höhe ($p < 0{,}01$ und r^2 indiziert eine mittlere bis hohe Effektstärke) erreichen.

Tabelle 2: Zusammenhänge zwischen MARKUS-V und MARKUS-C auf der Ebene der Bildungsgänge

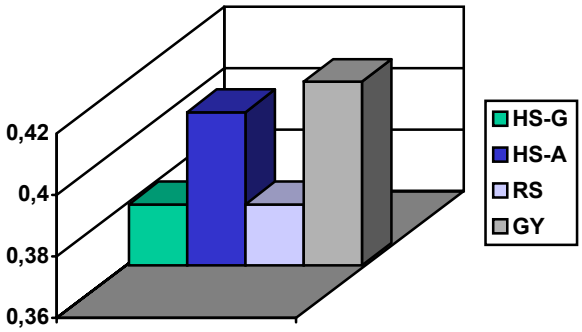
HS-G	**HS-A**	**RS**	**GY**
.39	.37	.32	.33

Dieser bildungsgangspezifischen Moderierung soll nunmehr auch im Kontext von MARKUS-T nachgegangen werden. Es interessiert dabei, wie die Validitätskoeffizienten in Abhängigkeit vom jeweiligen Bildungsgang variieren.

Die nachfolgende Abbildung 2 zeigt auf, in welche Richtung die Ergebnisse deuten. Die korrelativen Zusammenhänge werden nunmehr – wie in Tabelle 2 – kleiner. Hierbei sind die Zusammenhänge bei der Gruppe HS-A und GY größer als für die anderen Bildungsgänge. Inwieweit dieses Ergebnis durch die Tatsache moderiert wird, dass unterschiedliche Lerngelegenheiten gegeben waren, kann an

dieser Stelle nur hypothetisch unterstellt werden. Dieser Aspekt der Lerngelegenheiten wird in Helmke & Jäger (im Druck) ausführlich behandelt.

Abbildung 2: Bildungsgangspezifische Zusammenhänge zwischen MARKUS-C und MARKUS-T



Wenn solche vergleichsweise hohen korrelativen Übereinstimmungen vorhanden sind, so lohnt es sich der Frage nachzugehen, ob und in welchem Ausmaß aus der Kenntnis von $MARKUS_{Fundamentum}$ und $MARKUS_{Additum}$ die Ergebnisse von MARKUS-T rekonstruiert werden können.

Dieser Fragestellung (s. Fragestellung b) soll anhand einer linearen Regression nachgegangen werden. Hierbei wird von folgendem linearen Modell ausgegangen:

Formel 2:

$$\mathbf{MARKUS\text{-}T' = b1 * MARKUS_{Fundamentum} + b2 * MARKUS_{Additum} + a}$$

Allerdings interessiert dabei weniger der Grad des Zusammenhangs (dargestellt durch R^2), sondern wie gut die Rekonstruktionen gelingen. Als Maß hierfür werden die Residualwerte herangezogen:

Formel 3:

$$\mathbf{D = MARKUS\text{-}T' - MARKUS\text{-}T}$$

Um die Berechnung zu vereinfachen, wurden alle in den Formeln 1, 2 und 3 verwendeten Werte z-transformiert. Da gemäß Formel 3 kontinuierliche Werte resultieren, soll unter mehr pragmatischem Blickwinkel eine Art von Normierung erfolgen. Diese wird entsprechend der nachfolgend dargestellten Tabelle 3 wie folgt realisiert: Analog der zweiten Zeile von Tabelle 3 werden auf der Grundlage der z-transformierten Residualwerte Klassifikationen vorgenommen. Diese Klassifikationen erlauben dann der Frage nachzugehen, ob und in welchem Ausmaß Über- bzw. Unterschätzungen in bestimmten Bildungsgängen erzeugt werden.

Tabelle 3: Über- und Unterschätzungen von MARKUS-T auf der Basis von MARKUS-C

	1	**2**	**3**	**4**	**5**	**6**
	$z \geq 2{,}0$	$1 \leq z < 2$	$0 \leq z < 1$	$0 > z > -1$	$-1 \geq z > -2$	$z \leq -2$
HS-G	**1,78%**	**14,51%**	53,17%	28,51%	1,97%	0,06%
HS-A	0,44%	6,37%	41,18%	44,16%	**6,83%**	**1,03%**
RS	0,33%	6,83%	42,93%	40,51%	**8,17%**	**1,23%**
GY	0,37%	4,52%	34,49%	43,03%	**14,32%**	**32,27%**

Tabelle 3 ist nun so zu interpretieren, dass insbesondere bei der Teilgruppe HS-G Überschätzungen auftreten. In diesem Falle wird gemäß Formel 3 der Residualwert positiv. Im Vergleich der beiden Klassifikationsstufen 1 und 2 gegenüber 5 und 6 ist der prozentuale Anteil der Überschätzungen in der Teilgruppe HS-G signifikant größer ($p < 0{,}01$) als die der Unterschätzungen. Umgekehrt aber werden auf der Basis einer für alle Schülerinnen und Schüler vorgenommenen Schätzung die Werte der Schülerinnen und Schüler aus den Bildungsgängen HS-A, RS und GY eher unterschätzt. Der Anteil von Unterschätzungen ist dabei für HS-A, RS und GY signifikant größer als der der Überschätzungen ($p < 0{,}01$). Wiederum zeigt sich ein differenzieller Effekt unter Zugrundelegung des Bildungsgangs.

Im Folgenden soll der Frage nachgegangen werden, wie die Personengruppe beschrieben werden kann, die in HS-G als Überschätzte gelten. In einem ersten Schritt fallen hierzu folgende Sachverhalte auf:

Greift man nunmehr nur die Stufe 1 aus Tabelle 3 heraus, so stellt man Folgendes fest: Von den insgesamt 213 Personen wird als Sprache, die zuhause gesprochen wird, Deutsch zu 74,1%, Russisch zu 9,6% und Türkisch zu 6,1% angegeben. Alle anderen Sprachgruppen sind nur marginal vertreten. Die Gruppe der überschätzten Personen wird also in der Mehrheit nicht auf Grund einer anderen Sprache als Deutsch definiert.

Gibt es in einem zweiten Schritt Hinweise dafür, dass diese Gruppe aus bestimmten Klassen mit bestimmten Merkmalen stammt? Ein Hinweis ergibt sich aus Tabelle 4: Hier werden diejenigen Klassen beschrieben, welche 25% und mehr Überschätzte nach dem oben genannten Kriterium enthalten. Bei insgesamt N = 54 Klassen ergibt sich hinsichtlich der Leistungsfähigkeit der Klasse (eingeschätzt durch den Mathematiklehrer) ein Notenmittel (arithmetisches Mittel) von 4,48 für die betreffenden Klassen sowie einer Standardabweichung SD = 0,41 (vgl. Tabelle 4).

Tabelle 4: Verteilungsparameter von Klassen aus der Gruppe der Überschätzten

Variable	N	Mittelw.	Min.	Max.	SD
Mathematiknote (Halbjahreszeugnis)	54	4,48	3,43	5,25	,41

Mit diesen Daten verdichtet sich das Indiz, dass es sich um solche Schülerinnen und Schüler handelt, welche eher am unteren Ende der Leistungsverteilung angesiedelt sind. Zugleich handelt es sich um solche Personen, die in einem Umfeld unterrichtet werden, dessen Leistungsniveau ebenso als unterdurchschnittlich eingeschätzt wird.

Dieses Indiz wird durch folgenden Sachverhalt erhärtet: Stellt man die Gruppe der Über- und Unterschätzten aus HS-G gegenüber (das sind diejenigen, welche den Klassifikationsgruppen 1 und 2 bzw. 5 und 6 aus Tabelle 3 entsprechen), so zeigt sich bei einer Gegenüberstellung mit Hilfe eines t-Test für unabhängige Stichproben mit Blick auf das Vorwissen (erfasst durch MARKUC-V), Noten in Mathematik 1. Halbjahr und 2. Halbjahr sowie der Note, die sich ein Schüler selbst zuschreibt, wenn er sich sehr anstrengen würde, folgendes Resultat (s. Tabelle 5):

1. Beide Teilgruppen unterscheiden sich nicht auf der Basis der verschiedenen Mathematiknoten (Halbjahreszeugnis, erwartete Note und unter Anstrengung zu erreichende Mathematiknote): $p > 0,05$.

2. Hingegen resultiert hinsichtlich des Vorwissens ein signifikanter Unterschied zwischen den Gruppen ($p < 0{,}01$). Die dazugehörige Effektgröße ω^2 von .30 spricht für einen großen Effekt.

3. Die regressionsanalytisch identifizierten überschätzten Schülerinnen und Schüler können damit als solche erkannt werden, welche zum einen aus einem Klassenumfeld stammen, das im unteren Leistungsniveau angesiedelt ist (s. Tabelle 4), und zum anderen mit einem vergleichsweise geringen Vorwissen ausgestattet sind.

Tabelle 5: Mittelwertsvergleich zwischen Unter- und Überschätzten aus der HS-G

	Mittelw. Überschätzte	Mittelw. Unterschätzte	t-Wert	FG	P
MARKUS-V	-1,48	,183	-27,85	1854	0,00
Mathematiknote (Halbjahreszeugnis)	3,48	3,53	-,71	1561	0,48
Mathematiknote (erwartet)	2,52	2,47	,71	1561	0,48
Mathematiknote (Anstrengung)	3,26	3,29	-,45	1573	0,65

Diskussion

Es war das Ziel dieses Beitrags, zwei Aspekte aufzuzeigen:

1. Wie auf der Grundlage einer Teilgruppe von Items aus der ursprünglichen Untersuchung von TIMSS die Leistungsfähigkeit von Schülerinnen und Schülern (hier aus Rheinland-Pfalz) in Mathematik eingeschätzt und mit bereits vorliegenden Ergebnissen verglichen werden kann. Hierbei zeigt sich, dass die in der MARKUS-Studie Untersuchten auf keinen Fall schlechter abgeschnitten haben als die ehemals untersuchte repräsentative Stichprobe bei TIMSS: Es spricht im Gegenteil Vieles dafür, dass die Leistungen dieser Schülerinnen und Schüler besser sind als in der TIMSS-Originalstichprobe aus der Bundesrepublik Deutschland. Da zum Zeitpunkt dieser Veröffentlichung noch keine Daten von der PISA-Studie über die Ländervergleiche aus der Bundesrepublik Deutschland vorliegen, kann dieses Ergebnis allerdings als eine Art von Prognosewert angesehen werden. Der Vergleich wird im Dezember 2002 möglich sein.

2. Wie verfahren werden kann, um das Problem der Validierung von MARKUS-C anzugehen. Hier wurde wiederum in TIMSS ein Ausgangspunkt gesucht. Dabei zeigte sich eine Reihe von differenziellen Befunden, welche nahe legt, dass die Konzepte von MARKUS und TIMSS in den je gewählten Ausschnitten nur in Teilen zur gegenseitigen Validierung herangezogen werden können. Ein wesentlicher Befund war hierbei, dass offensichtlich die TIMSS-Items zu überproportional guten Einschätzungen der Leistungsfähigkeit führen. Diese Prognose wurde genutzt, um einen Teil der Stichprobe besser identifizieren zu können: Es handelt sich hierbei um Schülerinnen und Schüler aus der HS-G, welche sowohl über ein schlechtes Vorwissen verfügen als auch in einem leistungsmäßig unteren Klassenumfeld angesiedelt sind. Solche Teilgruppen zu identifizieren, muss eine besondere Aufgabe der Bildungsforschung sein. Wenn man die Ergebnisse von PISA (Baumert et al., 2001) konsequent für eine Umorientierung der pädagogischen Praxis umsetzen will, so ist auch in diagnostischer Sicht ein leistungsfähiges Instrument notwendig, das zur Identifikation von Teilgruppen eingesetzt werden muss, um mit diesem förderungsbedürftige Schülerinnen und Schüler bereits auf der Ebene der einzelnen Klasse auszumachen. Hierzu hat die MARKUS-Studie einen Beitrag geleistet. Die in diesem Beitrag berichteten Daten belegen dies.

Literatur

Balzer, L. & Jäger, R. S. (2001). Fachleistung Mathematik in MARKUS . Empirische Pädagogik, 15, 4, S. 535-551.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M; Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J. & Weiß, M. (Hrsg.). (2001). PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich.

Baumert, J., Bos, W. & R. Lehmann (Hrsg.) (2000). TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn (Bd. 1). Opladen: Leske + Budrich.

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller, O. & Neubrand, J. (1997). TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde. Opladen: Leske + Budrich.

Bortz, J. & Döring, N. (1995). Forschungsmethoden und Evaluation (2. Aufl.). Berlin: Springer.

Helmke, A., Jäger, R. S., Balzer, L., Hosenfeld, I., Ridder, A. & Schrader, (2001). Das Projekt MARKUS: Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext. In G. Kaiser, N. Knoche, D. Lind & W. Zillmer (Hrsg.). Leistungsvergleiche im Mathematikunterricht – ein Überblick über aktuelle nationale Studien (S. 51-93). Hildesheim: Verlag Franzbecker.

Helmke, A. & Jäger, R. S. (Hrsg.). (im Druck). MARKUS - Mathematik Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext. Landau: Verlag Empirische Pädagogik.

Hornke, L. F. & Habon, M. (1984). Erfahrungen zur rationalen Konstruktion von Testaufgaben. Zeitschrift für Differentielle und Diagnostische Psychologie, 5, 203-212.

Jäger, R. S. (2001). Von der Beobachtung zur Notengebung (4. vollst. überarb. Auflage). Landau: Verlag Empirische Pädagogik.

Klieme, E., Baumert, J., Köller, O. & Bos, W. (1997). In Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller O. & Neubrand, J. (1997). TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde (S. 85-133). Opladen: Leske + Budrich.

Lehmann, R. H., Peeck, R., Gänsfuß, R., Lutkat, S., Mücke, S. & Barth, I. (2000). QuaSUM. Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik. Ergebnisse einer repräsentativen Untersuchung im Land Brandenburg. Potsdam: Ministerium für Bildung, Jugend und Sport des Landes Brandenburg.

Anschrift der Autoren:

Prof. Dr. Reinhold S. Jäger, Zentrum für empirische pädagogische Forschung (ZepF) der Universität Koblenz-Landau, Campus Landau, Friedrich-Ebert-Str. 12, D-76829 Landau. Email: jaeger@zepf.uni-landau.de

Lars Balzer, Dipl.-Psych., Zentrum für empirische pädagogische Forschung (ZepF) der Universität Koblenz-Landau, Campus Landau, Friedrich-Ebert-Str. 12, D-76829 Landau. Email: balzer@zepf.uni-landau.de

Jäger, R. S. & Balzer. L. (2001). MARKUS und TIMSS – ein Vergleich. Empirische Pädagogik, 15, 553-566

Auf der Grundlage einer der größten Mathematik-Vergleichsuntersuchungen in der Bundesrepublik Deutschland – MARKUS – werden einige Teilfragen angegangen, welche von allgemeinem Interesse sind. Es sind dies: (1) Welches Niveau hätten die Schülerinnen und Schüler aus Rheinland-Pfalz erreicht, wenn sie an TIMSS teilgenommen hätten? (2) Wie gut können die Leistungen von TIMSS vorhergesagt werden, wenn der Ausgangpunkt bei MARKUS liegt? Zur Beantwortung dieser Fragen werden unterschiedliche Methoden angewendet.

Jäger, R. S. & Balzer. L. (2001). MARKUS and TIMSS – a comparison. Empirische Pädagogik, 15, 553-566

On the basis of a large scale assessment (MARKUS) in Rhein-Palatine several questions are tested: (1) What would be the level of 8th grade students if they participated in the TIMSS, (2) Can TIMSS measures be predicted with regressions methods knowing the measures of the two main parts of the MARKUS-test? Answering these main questions several methods and data are reported.