Zeitschrift für Erziehungswissenschaft (2023) 26:1321–1347
https://doi.org/10.1007/s11618-023-01176-6

ZfE

ALLGEMEINER TEIL

Früher war alles besser? Mathematikleistungen von Abiturientinnen und Abiturienten von 1964 und 1996 im Vergleich

Tobias Rolfes · Alexander Robitzsch · Aiso Heinze

Eingegangen: 22. Oktober 2021 / Überarbeitet: 30. März 2023 / Angenommen: 16. Mai 2023 / Online publiziert: 8. August 2023

Zusammenfassung Gemäß der Wahrnehmung insbesondere von Hochschullehrenden verringern sich die Fähigkeiten der Abiturientinnen und Abiturienten im Fach Mathematik seit Jahrzehnten beständig. Allerdings liegen bisher kaum empirische Untersuchungen zur Trendentwicklung der Mathematikleistungen in der gymnasialen Oberstufe vor. Um der Frage nachzugehen, ob sich die vermutete negative Trendentwicklung empirisch nachweisen lässt, wurden die Mathematikleistungen von Abiturienten und Abiturienten in Hessen und Schleswig-Holstein untersucht. Dazu wurde eine Sekundäranalyse der Daten aus der First International Mathematics Study (FIMS) von 1964 und der Third International Science and Mathematics Study (TIMSS) von 1996 vorgenommen. Dabei wurden die Daten aus FIMS und TIMSS mit Hilfe der Item-Response-Theorie neu skaliert und anhand der neun Trenditems über ein Mean-Mean-Linking verbunden. Anschließend wurden die Mathematikleistungen von 1964 und 1996 durch ein Equipercentile-Equating in die TIMSS-Metrik überführt und in das TIMSS-Kompetenzstufenmodell eingeordnet. Die Ergebnisse zeigten, dass sich die Mathematikleistungen der Abiturientinnen und Abiturienten von 1964 und 1996 in den beiden Bundesländern Hessen und Schleswig-Holstein

✉ Prof. Dr. Tobias Rolfes
Institut für Didaktik der Mathematik und der Informatik, Goethe-Universität Frankfurt am Main, Robert-Mayer-Str. 6–8, 60325 Frankfurt am Main, Deutschland
E-Mail: rolfes@math.uni-frankfurt.de

Prof. Dr. Tobias Rolfes · Dr. Alexander Robitzsch · Prof. Dr. Aiso Heinze
IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Olshausenstraße 62, 24118 Kiel, Deutschland

Dr. Alexander Robitzsch
E-Mail: robitzsch@leibniz-ipn.de

Prof. Dr. Aiso Heinze
E-Mail: heinze@leibniz-ipn.de

nicht signifikant unterschieden und sich die vermutete negative Trendentwicklung für diese beiden Bundesländer empirisch nicht belegen lässt.

Schlüsselwörter FIMS · TIMSS · Gymnasiale Oberstufe · Mathematikleistungen · Trendentwicklung · Sekundäranalysen

Everything was better in the past? Comparing mathematics performance of upper secondary school graduates in 1964 and 1996

Abstract According to the perception in particular of universities teachers, the skills of upper secondary school graduates in mathematics have been steadily declining for decades. However, there have been hardly any empirical studies on the trend development of mathematics performance in upper secondary schools. In order to investigate whether the assumed negative trend development can be empirically proven, the mathematics performance of upper secondary school graduates in Hesse and Schleswig-Holstein was examined. For this purpose, we conducted a secondary analysis of data from the First International Mathematics Study (FIMS) of 1964 and the Third International Science and Mathematics Study (TIMSS) of 1996. In this analysis, the data from FIMS and TIMSS were rescaled using item response theory and linked using the nine trend items via mean-mean linkage. Subsequently, the 1964 and 1996 mathematics performances were transformed into the TIMSS metric by an equipercentile equating and classified into the TIMSS proficiency level model. The results showed no significant difference in mathematics performance between upper secondary school graduates of 1964 and 1996 in Hesse and Schleswig-Holstein and that the assumed negative trend development could not be empirically substantiated for these two federal states.

Keywords FIMS · TIMSS · Upper secondary school · Mathematics performance · Trend development · Secondary analyses

1 Einleitung

Die empirische Untersuchung der schulischen Leistungen von Schülerinnen und Schülern ist seit der Jahrtausendwende verstärkt in den Blick der Bildungsforschung gerückt. Einen zentralen Ausgangspunkt dieser „empirischen Wende“ bildete die *Third International Mathematics and Science Study* (TIMSS) aus den Jahren 1995 und 1996, bei der sich die Leistungen der Schülerinnen und Schüler aus Deutschland im internationalen Vergleich nur als mittelmäßig erwiesen (Baumert et al. 2000a, b; Baumert und Lehmann 1997). Als eine Konsequenz aus den TIMSS-Ergebnissen wurde von der Bildungspolitik die regelmäßige Teilnahme an internationalen Schulleistungsstudien (z. B. PISA, TIMSS und PIRLS/IGLU) und schließlich die zentrale Überprüfung des Erreichens der Bildungsstandards im Ländervergleich (IQB-Bildungstrend) beschlossen.

Allerdings konzentrieren sich diese empirischen Untersuchungen bisher vornehmlich auf die Primarstufe und die Sekundarstufe I. Zwar ist die gymnasiale Oberstufe

beständig Gegenstand der bildungspolitischen Auseinandersetzung, in der zum Beispiel um die Ausgestaltung der Oberstufe (Kurssystem versus Profiloberstufe) oder der Abiturprüfungen (zentral versus dezentral) debattiert wird. Ein Bildungsmonitoring des Leistungsstandes der Abiturientinnen und Abiturienten findet aktuell jedoch nicht statt und ist auch in naher Zukunft nicht geplant (Neumann und Trautwein 2019; Stanat et al. 2016). Da mittlerweile bundesweit durchschnittlich 40 % eines Altersjahrgangs das Abitur ablegen (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK] 2019) und somit die gymnasiale Oberstufe über die letzten Jahrzehnte einen immer größeren Anteil des Bildungssystems ausmacht, ist diese mangelnde empirische Datenlage zur gymnasialen Oberstufe problematisch. Dementsprechend nehmen die TIMS-Studien aus 1995/1996 eine Sonderstellung ein, da Deutschland an dieser Schulleistungsuntersuchung auch mit den Abschlussklassen der Sekundarstufe II teilnahm.

Insbesondere beim Übergang von der Schule zur Hochschule sind die Mathematikleistungen der Abiturientinnen und Abiturienten ein zentrales Thema. Hierbei haben die Klagen über das geringe mathematische Leistungsniveau der Schulabgängerinnen und Schulabgänger eine gewisse Tradition. Bereits in den 1950er-Jahren wurden von den Hochschulen schlechte Mathematikkenntnisse der Studienanfängerinnen und Studienanfänger beklagt (Behnke 1965; Steiner 1984). Auch in den 1980er-Jahren kritisierten mathematisch-naturwissenschaftliche Fachverbände bei Abiturientinnen und Abiturienten „einen Verfall der Kenntnisse und Fähigkeiten in den mathematischen, naturwissenschaftlichen und technischen Fächern“ (DMV et al. 1982, S. 25). Ebenso wurde in den 1990er-Jahren bei Studienanfängerinnen und Studienanfängern die „Sicherheit im Umgang mit mathematischen Symbolen und Modellen“ (KMK 1995, S. 24) vermisst. Bis heute hat die Kritik an den Mathematikleistungen der Abiturientinnen und Abiturienten nicht abgenommen. Auch 2017 stellten circa 130 Hochschuldozierenden aus dem Fach Mathematik in einem offenen Brief fest, dass „das mathematische Vorwissen von vielen Studienanfängern nicht mehr für ein WiMINT-Studium ausreicht“ („Mathematikunterricht und Kompetenzorientierung – ein offener Brief“ 2017, S. 1).

Angesichts dieser von Hochschulseite wahrgenommenen negativen Trendentwicklung im Leistungsniveau in Mathematik stellt sich die Frage, inwieweit es empirische Evidenz für diese Einschätzung gibt oder ob diese Einschätzungen einer verzerrten Wahrnehmung entspringen. So kann beim Menschen die Trendwahrnehmung vom Phänomen der Nostalgie (Batcho 1995; Leboe und Ansons 2006) geprägt sein, welches zuweilen mit dem Bonmot „Nostalgia is like a grammar lesson: you find the present tense and the past perfect“ charakterisiert wird. Beispielsweise zeigte sich, dass der Trend bezüglich globaler Themen wie Wohlstand, Gesundheit und Kriminalität von Menschen negativer eingeschätzt wird, als er in Wirklichkeit ist (Pinker 2018; Rosling et al. 2018). Aus den Erziehungswissenschaften bzw. der Bildungsforschung sind keine Untersuchungen zu Erinnerungseffekten und ihren Auswirkungen bekannt. Es ist aber anzunehmen, dass die Erinnerung der Vergangenheit auch im Bildungsbereich einen bedeutenden Einfluss hat. Beispielsweise wird in dem oben genannten offenen Brief von 2017 eine Abkehr von den Bildungsstandards und der Kompetenzorientierung und eine stärkere Orientierung an curricularen Inhalten vor der Zeit der Bildungsstandards gefordert. Diese Schlussfol-

gerung beruht auf der Prämisse, dass die Mathematikleistungen der Abiturientinnen und Abiturienten „früher“ besser waren. Somit ist die Frage, ob sich die Mathematikleistungen der Abiturientinnen und Abiturienten verschlechtert haben oder ob diese Wahrnehmung möglicherweise durch Nostalgie verzerrt ist, nicht nur eine akademische Frage, sondern hat auch Auswirkungen auf den bildungspolitischen Diskurs über die Ausgestaltung des Mathematikunterrichts in der gymnasialen Oberstufe.

Eine Annäherung an die Frage, ob empirische Evidenz für die vermutete negative Trendentwicklung bei den Mathematikleistungen in der gymnasialen Oberstufe generiert werden kann, eröffnet der Vergleich von Mathematikleistungen von Abiturientinnen und Abiturienten der Jahre 1964 und 1996. Wie bereits dargestellt, wurden im Rahmen der bundesweiten TIMS-Studie im Jahre 1996 die Mathematikleistungen des Abschlussjahrgangs der Sekundarstufe II evaluiert. Aus dem Jahr 1964 liegen Daten der *First International Mathematics Study* (FIMS) vor, die in den Bundesländern Hessen und Schleswig-Holstein die Mathematikleistungen der Abiturientinnen und Abiturienten untersuchte (Husén 1967; Schultze und Riemenschneider 1967). Die Daten von FIMS wurden damals nur eingeschränkt ausgewertet und sind daher in Deutschland bisher kaum rezipiert worden. Auch international sind Trenduntersuchungen unter Einbezug der FIMS-Daten bisher nur für die Population aus der Sekundarstufe I durchgeführt worden (Majoros et al. 2021).

Wie eine von uns durchgeführte Itemanalyse zeigte, wurden zehn Items aus FIMS auch in TIMSS verwendet, sodass ein statistischer Vergleich der Mathematikleistungen zu den beiden Zeitpunkten möglich ist. Da kaum empirische Daten über die Mathematikleistungen in der frühen Bundesrepublik existieren und die FIMS-Daten bisher kaum ausgewertet wurden, kann der Vergleich der Mathematikleistungen der Abiturientinnen und Abiturienten von 1964 und 1996 interessante Einblicke in die Trendentwicklung geben, die nicht nur einen dokumentarischen Wert besitzen, sondern auch Implikationen bezüglich der organisatorischen und inhaltlichen Ausgestaltung des Oberstufenunterrichts in Mathematik haben könnten.

Daher berichtet der vorliegende Artikel eine Sekundäranalyse der FIMS- und TIMSS-Daten, bei der die Mathematikleistungen der Abiturientinnen und Abiturienten von 1964 und 1996 verglichen werden. Schwerpunktmäßig wird dabei der Frage nachgegangen, ob sich die vermutete negative Trendentwicklung in dem 32-jährigen Zeitraum von FIMS zu TIMSS empirisch belegen lässt. Im Theorieteil wird dazu zunächst reflektiert, welche empirischen Befunde es zu den Mathematikleistungen in der gymnasialen Oberstufe und ihre Trendentwicklung in und außerhalb Deutschlands gibt. In einem zweiten Teil wird gegenübergestellt, wie der Mathematikunterricht in der gymnasialen Oberstufe in den Jahren 1964 und 1996 organisatorisch und inhaltlich ausgestaltet war.

2 Theoretischer Hintergrund

2.1 Empirische Befunde über die Mathematikleistungen und ihre Trendentwicklung

Bezüglich der Mathematikleistungen in der Sekundarstufe I zeigte sich in einer Trendstudie, dass die Mathematikleistungen von Siebtklässlern an nordrhein-westfälischen Gymnasien im Zeitraum von 1969 bis 1997 statistisch signifikant um $d = 0{,}32$ abnahmen (Becker et al. 2006). Vergleichbare Daten liegen für die Sekundarstufe II bisher nicht vor, da die Datenlage bezüglich der Mathematikleistungen der Abiturientinnen und Abiturienten und ihrer Trendentwicklung in Deutschland sehr limitiert ist. Gleichwohl hat es in den letzten 25 Jahren seit TIMSS einige Schulleistungsuntersuchungen in einzelnen Bundesländern gegeben, die Grundlage einer Reanalyse (Rolfes et al. 2021) waren. Dazu wurde das Kompetenzstufenmodell nach Klieme (2000) für den TIMSS-Mathematiktest zur voruniversitären Mathematik herangezogen, um die Ergebnisse aus den unterschiedlichen Schulleistungsstudien kriterial einordnen zu können. Hierbei zeigte sich, dass es deutliche Leistungsdisparitäten zwischen Bundesländern gab. Aber selbst in leistungsstarken Bundesländern konnten deutliche Leistungsdefizite bei den Mathematikleistungen festgestellt werden, da auch dort nur eine Minderheit der Abiturientinnen und Abiturienten das definierte Mindestniveau des Kompetenzstufenmodells nach Klieme (2000) erreichte. Eine eindeutig negative Trendentwicklung der Mathematikleistungen seit 1996 war insgesamt nicht feststellbar, auch wenn die Datenlage hierzu sehr begrenzt war (Rolfes et al. 2021).

Eine weitere empirische Studie, die Rückschlüsse auf die Trendentwicklung der Mathematikleistungen von Abiturientinnen und Abiturienten erlaubt, ist eine von Buschhüter et al. (2016) durchgeführte Untersuchung in der Studieneingangsphase. Hierbei wurde im Wintersemester 2013/14 ein 1978 bundesweit eingesetzter Studieneingangstest erneut bundesweit mit 2322 Studienanfängerinnen und Studienanfängern in Physik durchgeführt. Die Ergebnisse ergaben keine empirische Evidenz dafür, dass die Testpersonen aus 2013 pauschal niedrigere Mathematikleistungen zeigten als die Testpersonen aus 1978. Dabei ist allerdings unklar, inwieweit sich diese Ergebnisse auf die Gesamtpopulation der Abiturientinnen und Abiturienten verallgemeinern lassen, da die Studienanfängerinnen und Studienanfänger des Faches Physik nur eine spezifische Teilstichprobe der Abiturientinnen und Abiturienten darstellen.

Einen Einblick in die Trendentwicklung der Mathematikleistungen in der Sekundarstufe II im internationalen Bereich erlauben die Schulleistungsstudien der *International Association for the Evaluation of Educational Achievement* (IEA). Deutschland hatte mit der Sekundarstufe II nur an der FIMS-Erhebung im Jahr 1964 und an der TIMSS-Erhebung im Jahre 1995/96 teilgenommen. Neben FIMS und TIMSS administrierte das IEA für die Sekundarstufe II noch in den Jahren 1980–82 die *Second International Mathematics Study* (SIMS) und in den Jahren 2008 und 2015 in Fortführung von TIMSS die sogenannten *TIMSS Advanced*-Studien. Dabei zeigten Trendanalysen ein heterogenes Bild. In Schottland, Israel, England/Wales und Belgien zeigte sich deskriptiv eine negative Trendentwicklung von FIMS zu SIMS

(Robitaille und Taylor 1989). In Schweden, den Vereinigten Staaten, Finnland und Japan war in diesem Zeitraum dagegen ein positiver Trend zu erkennen. Ein ähnlich heterogenes Bild zeigte sich von TIMSS 1995 zu TIMSS Advanced 2015 (Mullis et al. 2016). In Frankreich, Italien und Schweden nahmen die Mathematikleistungen signifikant ab, während sich in den Vereinigten Staaten, Slowenien und Russland keine signifikante Leistungsänderung innerhalb dieser 20 Jahre zeigte. Bei den Trendschätzungen von TIMSS Advanced 2008 zu 2015 konnte im Libanon eine signifikante Abnahme und in Schweden eine signifikante Zunahme der Testleistungen identifiziert werden. Insgesamt verdeutlichen die Ergebnisse der IAE-Studien, dass sich die Trendentwicklung sehr stark länderspezifisch ausgestaltete und international keine Tendenz in der Trendentwicklung für die Mathematikleistungen in der Sekundarstufe II festgestellt werden kann.

Einen weiteren Einblick in die Trendentwicklung der Mathematikleistungen in der Sekundarstufe II über einen längeren Zeitraum erlauben Daten des *NAEP Long-Term Trend Assessment* (NAEP-LTT) in den Vereinigten Staaten. Bei NAEP-LTT wurde im Zeitraum von 1973 bis 2012 regelmäßig eine repräsentative Stichprobe aus Schülerinnen und Schüler aus öffentlichen und privaten Schulen befragt (National Center for Education Statistics 2012). Hierbei zeigte sich, dass die Leistungen der 17-Jährigen in diesem Zeitraum von 39 Jahren nahezu konstant blieben, während die Leistungen der 13-Jährigen und der 9-Jährigen zunahmen. Die Ergebnisse des NAEP-LTT verdeutlichen zum einen, dass ein Land über einen längeren Zeitraum stabile Leistungen in Mathematik in der Sekundarstufe II aufweisen kann. Zum anderen zeigen die Ergebnisse, dass die Trendentwicklung in der Sekundarstufe II in einem Land nicht parallel zum Trend in der Sekundarstufe I und Primarstufe verlaufen muss und in einem Bildungssystem ein separates Bildungsmonitoring der Sekundarstufe II durchaus sinnvoll ist.

Insgesamt kann daher festgestellt werden, dass

1. die Trendentwicklungen international zwischen Ländern differieren,
2. weder in Deutschland noch international bisher eine eindeutig negative Trendentwicklung bei den Mathematikleistungen in der Sekundarstufe II festzustellen ist,
3. über die Trendentwicklung der Mathematikleistungen in der Sekundarstufe II in Deutschland vor 1996 bisher nahezu nichts bekannt ist.

2.2 Mathematik in der gymnasialen Oberstufe im Jahr 1964 und 1996 im Vergleich

Im Folgenden erfolgt ein kurzer Abriss über die organisatorische und inhaltliche Ausgestaltung des Mathematikunterrichts in der gymnasialen Oberstufe in den Jahren 1964 (FIMS) und 1996 (TIMSS). Hierbei wird insbesondere auf die Bundesländer Hessen und Schleswig-Holstein Bezug genommen, da diese beiden Bundesländer an FIMS teilnahmen und daher die Trendentwicklung der Mathematikleistungen in diesen beiden Bundesländern im Fokus steht.

Im Jahr 1964 waren die Gymnasien in den altsprachlichen, den neusprachlichen und den mathematisch-naturwissenschaftlichen Zweig unterteilt (Schultze 1968). Wie Dokumentationen des Statistischen Bundesamtes (1967) zu entnehmen ist, be-

suchten Stand Mai 1964 von den Schülerinnen und Schülern der 13. Jahrgangsstufe in Hessen und Schleswig-Holstein etwa 40 % einen mathematisch-naturwissenschaftlichen Zweig und ungefähr 60 % einen neu- oder altsprachlichen Zweig. Auch in der Oberstufe fand der Unterricht in Klassenverbänden statt, allerdings wurde Mathematikunterricht in den verschiedenen Zweigen in unterschiedlichem Umfang erteilt. In Hessen wurde das Fach Mathematik in den Gymnasien des mathematisch-naturwissenschaftlichen Zweigs mit vier Stunden (Hessisches Ministerium für Erziehung und Volksbildung [HMEV] 1956) und in Schleswig-Holstein mit fünf bis sechs Stunden (Kultusminister des Landes Schleswig-Holstein [KMSH] 1956) bis zur Jahrgangsstufe 13 unterrichtet. In den Gymnasien des alt- und neusprachlichen Zweigs besuchten die Schülerinnen und Schüler dagegen in beiden Bundesländern nur in der Jahrgangsstufe 11 und 12 einen dreistündigen Mathematikunterricht (HMEV 1956; KMSH 1956). Inhaltlich war der Mathematikunterricht zu Beginn der 60er-Jahren von einer „Aufgabendidaktik“ (Lenné 1975, S. 50) geprägt, in der die Einübung mathematischer Verfahren im Vordergrund stand und die sich durch „fachstrukturelle Abstinenz“ (Lenné 1975, S. 54) auszeichnete. Da Mathematik zu den Kernpflichtfächern gehörte, mussten die Abiturientinnen und Abiturienten aller drei Gymnasialzweige in diesem Fach eine schriftliche Abiturprüfung ablegen (Schultze 1968).

Inhaltlich beschäftigte sich 1964 der Mathematikunterricht aller Zweige schwerpunktmäßig mit der *Analysis* (Differentiation, Integration) und der *Analytischen Geometrie der Kegelschnitte* (Punkt, Gerade, Kreis, Parabel, Ellipse, Hyperbel), wie damalige Lehrpläne (HMEV 1957; KMSH 1955) und Schulbücher (Lambacher und Schweizer 1959; Schweizer und Lambacher 1959) zeigen. Die Reformen der „Neuen Mathematik“, in denen die mathematischen Strukturen ins Zentrum des Mathematikunterrichts rückten (Büchter und Henn 2015), hatten 1964 noch keinen Einfluss auf den Mathematikunterricht in der gymnasialen Oberstufe.

Insgesamt waren die 1960er-Jahre in der Bundesrepublik durch einen kritischen Blick auf das eigene Bildungswesen geprägt. Als „Bildungskatastrophe“ charakterisierte Picht (1964) den Zustand des deutschen Bildungssystems im internationalen Vergleich. Insbesondere die niedrige Abiturientenquote in Deutschland (damalige OECD-Prognose für Deutschland für 1970: 6,8 %) im internationalen Vergleich (z. B. OECD-Prognose für Frankreich für 1970: 19 %) und die daraus resultierende zu geringe Anzahl an Studierenden sei problematisch (Picht 1964). Die Forderungen nach einer Erhöhung der Abiturientenquote traf aber auf eine geteilte Meinung, wie ein Zitat des Hamburger Landesschulrates Ernst Matthewes verdeutlicht: „Eine Erhöhung der Abiturientenzahl bedeutet zwangsläufig, daß die Gymnasien ihre Anforderungen senken. Anders läßt sich die Zahl der Abiturienten nicht vermehren.“ („Die letzte Hürde“ 1964, S. 81).

Letztendlich fand auch in Deutschland eine deutliche Erhöhung der Abiturientenquote statt. Während im Jahr 1964 im Bundesdurchschnitt die Abiturientenquote 8,1 % betrug, legten im Jahr 1996 24,2 % eines Altersjahrgangs das Abitur ab (Köhler et al. 2014). Im Jahr 1996 basierte der Unterricht in der gymnasialen Oberstufe organisatorisch auf der sogenannten reformierten Oberstufe, die im Zuge der Neugestaltung der gymnasialen Oberstufe 1972 (KMK 1972) eingeführt wurde. Ein Charakteristikum der reformierten Oberstufe war, dass die Klassenverbände in der

Qualifikationsphase aufgelöst wurden und die Schülerinnen und Schüler eine individuelle Kurskombination wählten. Die meisten der gewählten Fächer wurden dabei in einem Grundkurs auf einem grundlegenden Niveau unterrichtet. Für zwei Fächer konnten die Schülerinnen und Schüler einen Unterricht in Leistungskursen mit erhöhtem Anspruchsniveau und erhöhter Stundenzahl wählen. Im Jahr 1996 belegten im Bundesdurchschnitt 34,3 % der Schülerinnen und Schüler aus der Jahrgangsstufe 13 einen Leistungskurs in Mathematik (Baumert und Köller 2000).

In Schleswig-Holstein betrug für den Abiturjahrgang 1996 der Stundenumfang in Mathematik in den Grundkursen 3 Wochenstunden und in den Leistungskursen 5 Wochenstunden (KMSH 1980). In Hessen wurden die Leistungskurse 5–6-stündig und die Grundkurse 2–3-stündig unterrichtet (Land Hessen 1982), wobei davon ausgegangen werden kann, dass Mathematik im Grundkurs in der Regel dreistündig unterrichtet wurde. Damit hatten die Leistungskurse in Mathematik einen vergleichbaren Stundenumfang wie der Mathematikunterricht an den mathematisch-naturwissenschaftlichen Zweigen, während der Stundenumfang in den Grundkursen dem der sprachlichen Zweige entsprach.

Inhaltlich umfasste der Mathematikunterricht für den Abiturjahrgang 1996 schwerpunktmäßig die drei Sachgebiete *Analysis, Lineare Algebra/Analytische Geometrie* und *Stochastik* (KMK 1989). Im Bereich der Analysis waren wie 1964 die Differentiation und die Integration zentrale Gegenstände des Unterrichts, wobei im Vergleich zu 1964 stärker auch die theoretischen Grundlagen (Grenzwerte, Stetigkeit und Differenzierbarkeit) unterrichtliche Beachtung fanden (KMK 1989). Der Bereich Lineare Algebra/Analytische Geometrie unterschied sich 1996 deutlich von den Geometrieinhalten von 1964, da nun der Vektorbegriff anstelle der Kegelschnitte den roten Faden des Mathematikunterrichts in diesem Inhaltsgebiet darstellte (KMK 1989). War Stochastik in den 1960er-Jahren noch kein Gegenstand des Mathematikunterrichts, so war dieser Inhaltsbereich gemäß den Einheitlichen Prüfungsanforderungen möglicher Prüfungsteil des Abiturs im Jahr 1996, wobei allerdings nur zwei der drei Teilgebiete verpflichtend Gegenstand der Abiturprüfungen sein mussten (KMK 1989). So war im für den Abiturjahrgang 1996 gültigen Lehrplan von Schleswig-Holstein das Themengebiet Stochastik nicht verpflichtend und konnte beispielsweise durch das Themengebiet komplexe Zahlen ersetzt werden (KMSH 1986), während allerdings der hessische Kursstrukturplan Stochastik als verbindlichen Unterrichtsinhalt und auch Prüfungsgegenstand für die Abiturprüfung vorschrieb (Hessisches Kultusministerium 1991). Daher ist nicht auszuschließen, dass sich das Ausmaß der unterrichtlichen Behandlung der Stochastik in den beiden Bundesländern unterschied.

2.3 Methodische Herausforderungen bei Trendschätzungen

Trendschätzungen von Schulleistungen sind mit besonderen methodischen Herausforderungen verbunden. Eine zentrale Frage dabei ist, in welcher Weise die Ergebnisse aus einer Trendanalyse generalisierbar sind und ob die Unsicherheit quantifiziert werden kann. In der Generalisierbarkeitstheorie (Brennan 2001) wurden dafür *Facetten* definiert, welche die Generalisierbarkeit von Ergebnissen beeinflussen. Eine Facette stellt die *Personenauswahl* dar, wenn bei einer Untersuchung

nur eine bestimmte Stichprobe (z. B. Testpersonen in TIMSS) aus der Grundgesamtheit (alle Schülerinnen und Schüler des Abschlussjahrgangs der gymnasialen Oberstufe in Deutschland) befragt wird. Dieser Schätzfehler, der durch die *Personenauswahl* verursacht wird, ist auch Gegenstand klassischer inferenzstatistischer Verfahren und wird häufig in Form des Standardfehlers (*SE*) quantifiziert. Allerdings werden die Stichproben in Schulleistungsstudien gewöhnlicherweise nicht als einfache Zufallsstichprobe gezogen, sondern es werden Lerngruppen gesampelt, sodass alle Schülerinnen und Schüler einer Lerngruppe in der Stichprobe enthalten sind. Bei dieser Sampling-Methode würde die Anwendung der elementaren Verfahren der Inferenzstatistik für das Abschätzen des Standardfehlers des Mittelwertes ($SE = \widehat{\sigma}_{\overline{x}} = \frac{s_x}{\sqrt{n}}$) den tatsächlichen Fehler unterschätzen. In diesem Fall kommen häufig spezielle Verfahren, wie zum Beispiel das Jackknife-Verfahren (Kolenikov 2010), zum Einsatz, um den Fehler durch die Personenauswahl adäquater zu schätzen.

Neben der Verallgemeinerung über die Personengruppe der Stichprobe hinaus wird in Trendschätzungen von Schulleistungen auch über die konkret administrierte Itemmenge hinaus verallgemeinert. Es wird angenommen, dass die administrierten Items repräsentativ für einen Inhalts- bzw. Kompetenzbereich (d. h. eine größere Itemmenge) stehen sollen und daher eine Zufallsauswahl aus einem „Universum" aller möglichen Items darstellt. Daher stellt die *Itemauswahl* (*item sampling*) eine weitere Facette im Sinne der Generalisierbarkeitstheorie dar und ist neben der Personenauswahl eine weitere Fehlerquelle der Trendergebnisse. Eine statistische Möglichkeit zum Abschätzen dieses Fehlers ist das *Balanced Half Sampling* (Kolenikov 2010), bei dem Tests der halben Länge gebildet werden, die strukturell ähnlich zu den originalen Tests sind (siehe auch Robitzsch 2021). Wie Sekundäranalysen von Schulleistungsuntersuchungen zeigen, ist der Standardfehler durch Itemsampling zuweilen deutlich größer als der Standardfehler durch Personensampling (Robitzsch et al. 2011).

Bezüglich der administrierten Items stellt sich neben der Itemauswahl eine weitere methodische Herausforderung. In IRT-Modellierungen wird häufig ein invariantes Itemfunktionieren über Messzeitpunkte angenommen (Rupp und Zumbo 2006). In realen Datensätzen wird diese Annahme aber häufig verletzt und es tritt ein differenzielles Itemfunktionieren (DIF) auf (vgl. Wu et al. 2016). DIF kann dadurch behandelt werden, dass die Daten zu einzelnen Messzeitpunkten separat skaliert werden (z. B. mit einem 1PL-Modell) und mit einem anschließenden Linking-Verfahren (Kolen und Brennan 2014) auf eine gemeinsame Metrik gebracht werden. Als Ergebnis des Linking-Verfahrens kann man DIF-Effekte für die Linkitems ermitteln. Die Variabilität der DIF-Effekte wird in sogenannten Linkingfehlern (Monseur und Berezner 2007; Robitzsch und Lüdtke 2019; Wu 2010) quantifiziert.

3 Forschungsfragen

Wie zuvor dargestellt, gibt es zur Trendentwicklung der Mathematikleistungen von Abiturientinnen und Abiturienten in Deutschland über einen längeren Zeitraum zwar Wahrnehmungen, die insbesondere von Hochschulseite immer wieder geäußert wer-

den, Studien dazu fehlen aber. In dieser Arbeit sollen deshalb folgende Forschungsfragen beantwortet werden:

1. Welche Mathematikleistungen erbrachten die Abiturientinnen und Abiturienten 1964 im Vergleich zu 1996?
2. In welcher Weise differenzierten sich die Mathematikleistungen 1964 nach gymnasialen Schulzweig (mathematisch-naturwissenschaftlicher vs. sprachlicher Zweig) im Vergleich zu den Kursformen (Leistungskurs und Grundkurs) im Jahre 1996?
3. In welchem Ausmaß wird Unsicherheit in den Trendschätzungen von 1964 bis 1996 durch die Personenauswahl und in welchem Ausmaß durch die Itemauswahl evoziert?
4. Gibt es itemspezifische Unterschiede in der Trendentwicklung, d. h. gibt es differentielles Itemfunktionieren in der Trendschätzung?

4 Methode

4.1 Stichprobe

Grundlage der Sekundäranalysen des vorliegenden Artikels bildeten Daten der internationalen Schulleistungsstudien FIMS und TIMSS. Im Jahr 1964 fand FIMS (vgl. Husén 1967) in zwölf Ländern statt und war eine der ersten international durchgeführten empirischen Studien, um die Leistungsfähigkeit von Bildungssystemen zu vergleichen. Zielpopulation waren die 13-jährigen Schülerinnen und Schüler (Population 1) und der Abschlussjahrgang der hochschulvorbereitenden Sekundarstufe II (Population 3). Außerdem konnte optional ein national definierter Zwischenjahrgang (Population 2) teilnehmen. TIMSS war für das Erhebungsjahr 1995 angesetzt und fand in den 3. und 4. Klassen der Grundschule (Population I), den 7. und 8. Klassen der Sekundarstufe I (Population II) und in den Abschlussklassen der Sekundarstufe II (Population III) statt. Insgesamt nahmen 41 Länder an den TIMS-Studien teil.

4.1.1 FIMS

Aus FIMS bildeten die Daten[1] der Population 3 die Grundlage für unsere Sekundäranalysen. Die Population 3 wurde international altersunabhängig definiert und bestand aus allen Schülerinnen und Schülern des letzten Schuljahrs einer Vollzeitschule, von denen die Hochschulen oder äquivalente Institutionen der höheren Bildung ihre Studierenden rekrutieren (Husén 1967). Für Deutschland wurden die Schülerinnen und Schüler der Gymnasien im letzten Schuljahr unter die Population 3 gefasst (Schultze und Riemenschneider 1967). Dabei wurden als Population 3a der Abschlussjahrgang der Gymnasien des mathematisch-naturwissenschaftlichen

[1] Die FIMS-Daten wurden vom *Center for Comparative Analysis of Educational Achievement* (COMPEAT) der Universität Göteburg zu Verfügung gestellt (https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat/studies-before-1995/first-international-mathematics-study-1964).

Zweigs und als Population 3b der Abschlussjahrgang der Gymnasien des sprachlichen Zweigs (altsprachlicher und neusprachlicher Zweig) definiert (Schultze und Riemenschneider 1967).

Aus organisatorischen und finanziellen Gründen wurde FIMS in Deutschland nicht im gesamten Bundesgebiet durchgeführt, sondern die Ziehung einer repräsentativen Stichprobe beschränkte sich auf die Bundesländer Hessen und Schleswig-Holstein (Schultze und Riemenschneider 1967). Die Erhebung fand in der Zeit zwischen der schriftlichen und mündlichen Abiturprüfung des Jahres 1964 statt (Schultze und Riemenschneider 1967). Die beiden Unterpopulationen 3a und 3b waren in der Stichprobe nahezu in gleichem Umfang vertreten. Von der Population 3a wurden 649 Testpersonen (129 Schülerinnen und 520 Schüler, mittleres Alter: $M = 19{,}2$ Jahre, $SD = 3{,}8$ Jahre) aus 37 Gymnasien befragt. Aus der Population 3b nahmen 643 Testpersonen (288 Schülerinnen und 355 Schüler, mittleres Alter: $M = 18{,}3$ Jahre, $SD = 5{,}1$ Jahre) von 36 Gymnasien teil.

4.1.2 TIMSS

Aus TIMSS wurden Daten[2] einer Teilstichprobe der Population III verwendet. Nach internationaler Vereinbarung wurden als Population III die Schülerinnen und Schüler definiert, die sich zum Zeitpunkt der Erhebung im letzten Jahr der Sekundarstufe II in vollzeitlicher Ausbildung befanden (Martin und Kelly 1998). Innerhalb dieser Population III wurde eine Teilpopulation aus „students having taken advanced mathematics“ (Martin und Kelly 1998, S. 2) gebildet. In Deutschland wurde zu dieser Teilpopulation der gesamte Abschlussjahrgang einer gymnasialen Oberstufe gezählt, da in allen Bundesländern ein Besuch von voruniversitärem Mathematikunterricht für alle Schülerinnen und Schüler der gymnasialen Oberstufe mindestens bis zum Jahr vor dem Abschluss verpflichtend war (Baumert et al. 2000b).

Wegen organisatorischer Schwierigkeiten fand in Deutschland die Haupterhebung erst im Februar und März 1996 statt (Baumert et al. 2000b). Items zur voruniversitären Mathematik wurden bei 2246 Testpersonen (1412 Schülerinnen, 816 Schüler, 18 Personen ohne Angabe) aus 74 Schulen im gesamten Bundesgebiet administriert. Um eine Repräsentativität der Ergebnisse für die Bundesrepublik sicherzustellen, wurden von der nationalen TIMSS-Forschungsgruppe die Personengewichte für die deutsche Stichprobe adjustiert, da einzelne Bundesländer sowie Mathematik- und Physikleistungskurse in der Stichprobe überrepräsentiert waren (Baumert et al. 2000b). Das mittlere Alter der gewichteten Stichprobe betrug $M = 19{,}5$ Jahre ($SD = 0{,}8$ Jahre). Einen Leistungskurs in Mathematik besuchten 34 % der gewichteten Stichprobe, einen Grundkurs 66 %.

[2] Für die Analysen der TIMSS-Daten wurde von Rainer Watermann freundlicherweise der Datensatz, auf dem die Ergebnisse der nationalen Berichte für Deutschland (Baumert et al. 2000a, b) basierten, zur Verfügung gestellt.

4.2 Testmaterial, -design und -administration

In FIMS wurden in der Population 3 insgesamt 106 Items eingesetzt. Dabei wurden die Inhaltsgebiete *Arithmetik* (12 Items), *Algebra* (30 Items), *euklidische/analytische Geometrie* (23 Items), *Mengenlehre* (8 Items), *trigonometrische und Kreisfunktionen* (5 Items), *Analysis* (18 Items), *Wahrscheinlichkeitsrechnung* (2 Items) und *mathematische Logik* (8 Items) abgedeckt (Husén 1967). Die Items wurden auf sechs Testhefte, die keine gemeinsamen Items besaßen, verteilt. In der Population 3a bekamen die Testpersonen die vier Testhefte 5 (21 Items), 7 (17 Items), 8 (16 Items) und 9 (15 Items) vorgelegt, in der Population 3b die drei Testhefte 3 (20 Items), 5 (21 Items) und 6 (17 Items) (Husén 1967)[3]. Somit bildete das Testheft 5 mit 21 Items den Link zwischen den beiden Gymnasialzweigen. Für jedes Testheft wurde eine Bearbeitungszeit von 60 min eingeplant (Postlethwaite 1968), sodass den Testpersonen in der Population 3a insgesamt 240 min und in der Population 3b insgesamt 180 min für die Beantwortung der vier bzw. drei Testhefte zur Verfügung standen.

Bei TIMSS wurden insgesamt 68 Items in 65 Testlets aus fünf Inhaltsgebieten im Teilbereich voruniversitäre Mathematik eingesetzt (Baumert et al. 1999). Dabei wurden die Inhaltsgebiete *Zahlen, Gleichungen und Funktionen* (17 Items), *Analysis* (16 Items), *Geometrie/analytische Geometrie* (24 Items), *Wahrscheinlichkeitsrechnung/Statistik* (8 Items) und *Aussagenlogik/Beweise* (3 Items) abgedeckt.

Die Items zur voruniversitären Mathematik wurden auf vier Itemcluster (Cluster I: 10 Items, Cluster J: 21 Items, Cluster K: 18 Items, Cluster L: 19 Items) verteilt (vgl. Baumert et al. 2000b; Martin und Kelly 1998). In einem Multi-Matrix-Design wurden zur voruniversitären Mathematik drei verschiedene Testhefte (Testheft 3A: Cluster I und J, Testheft 3B: Cluster I und K, Testheft 3C: Cluster I und L) zusammengestellt. Außerdem enthielt ein gemischtes Testheft (4) das Cluster I gemeinsam mit Items zur mathematisch-naturwissenschaftlichen Grundbildung und zur voruniversitären Physik, um die drei unterschiedlichen bei TIMSS administrierten Teilbereiche zu verbinden. Die Bearbeitungszeit für ein Testheft betrug 90 min. Zehn Items aus FIMS wurden für die TIMS-Studien übernommen (vgl. Anhang). Bei einzelnen Items wurden in TIMSS leichte sprachliche Anpassungen (vor allem Anpassung der Übersetzung der im Englischen unveränderten Items) in den Instruktions- und Informationstexten vorgenommen. Der mathematische Inhalt und die Antwortmöglichkeiten waren bei diesen Trenditems jedoch identisch. Daher wird ein substanzieller Einfluss der leichten sprachlichen Variation auf die Itemschwierigkeiten nicht vermutet.

4.3 Datenanalyse

4.3.1 Skalierung der Leistungsdaten von FIMS und TIMSS

In den Primäranalysen wurde bei FIMS noch keine Item-Response-Theorie (IRT; Rost 2004) verwendet. Stattdessen wurden für die einzelnen Teilpopulationen 3a

[3] Die Testhefte 1, 2 und 4 wurden bei FIMS nur in der Erhebung in der Sekundarstufe I eingesetzt.

und 3b die durchschnittlichen Summenscores im jeweiligen Gesamttest für den Ländervergleich genutzt (Husén 1967; Schultze und Riemenschneider 1967). Dagegen wurden bei TIMSS in den Primäranalysen IRT-Modelle verwendet. Dabei wurden zunächst die Itemschwierigkeiten für die 68 Items anhand von Daten aus der internationalen Kalibrierungsstichprobe von 15.989 Testpersonen mit einem Partial-Credit-Modell geschätzt (vgl. Martin und Kelly 1998). In einem zweiten Schritt wurde das Modell für jedes Land separat angepasst, wobei Itemschwierigkeiten auf die Werte der internationalen Kalibrierungsstichprobe fixiert wurden. In Deutschland wurden für das Populationsmodell 60 Hintergrundvariablen (z. B. Geschlecht, Fähigkeitswert in Physik, Fähigkeitswert in mathematisch-naturwissenschaftlicher Grundbildung, Schulmittelwert in voruniversitärer Mathematik) einbezogen. Auf der Grundlage dieses Populationsmodells wurden 5 Plausible Values für die Personenfähigkeit in voruniversitärer Mathematik gezogen. Die Plausible Values wurden anschließend international in eine Skala mit einem Mittelwert von 500 und einer Standardabweichung von 100 transformiert (sog. 500/100-Skala).

In unserer Sekundäranalyse wurden für die Itemkalibrierung der FIMS- und TIMSS-Daten ausschließlich die nationalen Leistungsdaten aus Deutschland verwendet. Die Leistungsdaten von FIMS und TIMSS wurden separat jeweils mit einem Mehrgruppen-Partial-Credit-Modell skaliert. Bei FIMS bildeten die Gymnasialzweige (mathematisch-naturwissenschaftlich vs. sprachlich) und bei TIMSS die Kursform (Leistungskurs vs. Grundkurs) jeweils die beiden Gruppen. Bei FIMS wurden Personengewichte entsprechend den Anteilen der Gymnasialzweige (mathematisch-naturwissenschaftlich: 40 %; sprachlich: 60 %) generiert. Für die TIMSS-Stichprobe wurden die für Deutschland adjustierten Personengewichte des nationalen Berichts verwendet. Bezüglich der Eliminierung von Items aufgrund geringer Trennschärfe wurde eine verhältnismäßig liberale Schwelle von 0,15 festgelegt. Die in FIMS und TIMSS administrierten Items wurden jeweils von einer Expertengruppe entwickelt und ausgewählt und können somit eine hohe Konstruktvalidität beanspruchen. Eine solche deduktive Testkonstruktion auf der Basis von theoretischen Überlegungen wird als adäquates Vorgehen bei Kompetenzmessungen angesehen (vgl. Leutner et al. 2008). Deshalb sollte bei Kompetenzmessungen das Konstrukt auch nicht unnötig eingeschränkt werden, indem Items allein auf Grund von geringen Trennschärfen ausgeschlossen werden (Leutner et al. 2008). Nach dem Ausschluss der Items mit einer Trennschärfe unter 0,15 wurde zur Schätzung der Personenfähig-

Tab. 1 TIMSS-Kompetenzstufenmodell zur voruniversitären Mathematik (Klieme 2000, S. 87 ff.)

Scorewert	Niveau	Inhaltliche Charakterisierung
>600	Stufe IV	Selbstständiges Lösen mathematischer Probleme auf Oberstufenniveau
501–600	Stufe III (Mindestniveau Sek. II)	Anwendung von Lerninhalten der Oberstufe im Rahmen typischer Standardaufgaben
401–500	Stufe II	Anwendung einfacher mathematischer Begriffe und Regeln, die kein Verständnis von Konzepten der Oberstufenmathematik voraussetzen
≤400	Stufe I	Elementares Schlussfolgern

keiten der Testpersonen in beiden Populationen jeweils 20 Plausible Values gezogen, wobei kein Hintergrundmodell verwendet wurde.

4.3.2 Equating von FIMS und TIMSS

Das Linking der FIMS- und TIMSS-Population erfolgte durch ein Mean-Mean-Linking. Um die Mathematikleistungen auf der 500/100-Skala der TIMSS-Primäranalyse verorten zu können, wurde ein Equipercentile Equating zwischen den Personenfähigkeitswerten auf der 500/100-Skala der TIMSS-Primäranalyse und den Personenfähigkeitswerten unsere Sekundäranalyse der FIMS- und TIMSS-Daten durchgeführt. Zur kriterialen Einordnung der Personenfähigkeiten wurden die Fähigkeitswerte anschließend den Niveaustufen (vgl. Tab. 1) des TIMSS-Kompetenzstufenmodells für voruniversitäre Mathematik nach Klieme (2000) zugeordnet. Items der Niveaustufen I und II konnten noch mit Wissen und Fertigkeiten aus dem Mathematikunterricht der Sekundarstufe I gelöst werden und erst ab Stufe III waren Wissen und Fertigkeiten aus der gymnasialen Oberstufe erforderlich. Daher kann das Erreichen von Stufe III als Mindestniveau für Abiturientinnen und Abiturienten angesehen werden, da erst ab dieser Kompetenzstufe typische Aufgabenstellungen der Oberstufenmathematik erfolgreich bewältigt werden können (vgl. Klieme 2000).

4.3.3 Bestimmung statistischer Fehler und differenzielles Itemfunktionieren

Der Standardfehler *SE* für die Personenseite aufgrund der Stichprobenziehung von Personen wurde durch ein Jackknife-Verfahren geschätzt. Dabei wurden 34 Jackknifezonen mit jeweils 2 bis 4 Schulen gebildet (Kolenikov 2010). Es war dabei eine Abweichung vom originalen in TIMSS 1996 verwendeten Replikationsdesign notwendig, weil sich die Analysen unserer Studie nur auf die gymnasiale Teilpopulation der originalen Stichprobe bezogen. Schulen, die in TIMSS in der originalen Definition in einer Jackknifezone lagen, wurden auch in der Neudefinition in einer Jackknifezone zusammengefasst. Außerdem wurde in TIMSS das Stratum Bundesland für die Definition der Jackknifezonen berücksichtigt.

Zur Evaluierung der Unsicherheit bezüglich der Itemauswahl (Itemseite) wurde ein Linkingfehler *LE* (Gebhardt und Adams 2007) durch ein Balanced Half Sampling (Kolenikov 2010) bestimmt. Hierbei wurden jeweils Tests der halben Länge auf Basis von 32 Replikationszonen von Items gebildet, die strukturell ähnlich zu den originalen Tests sind. Der Gesamtfehler *TE* wurde definiert als $TE = \sqrt{SE^2 + LE^2}$. Abschließend wurde überprüft, in welchem Ausmaß ein differentielles Itemfunktionieren der Trenditems in der FIMS- und in der TIMSS-Population vorlag.

5 Ergebnisse

5.1 Skalierung FIMS/TIMSS und Linking

Wie im Methodenteil beschrieben, wurde für die Trennschärfe eine Schwelle von 0,15 festgelegt, d. h. alle Items mit Trennschärfen unterhalb dieses Wertes wurden

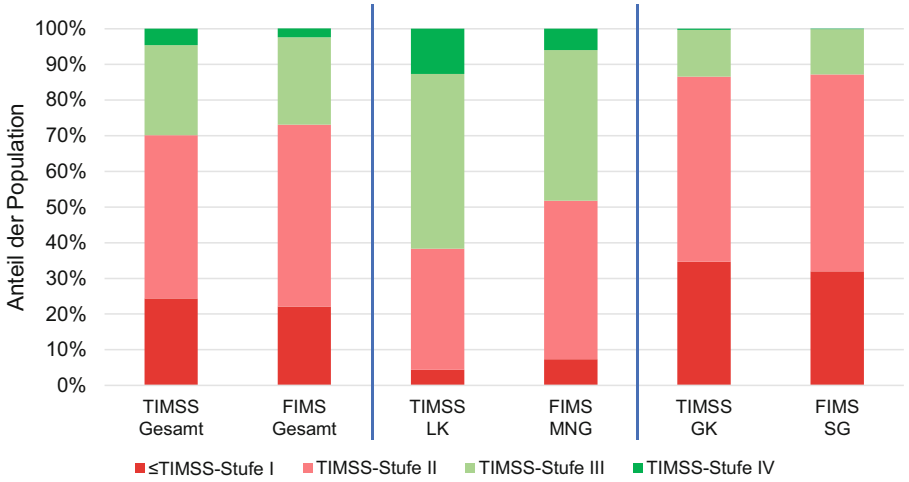

Abb. 1 Fähigkeitsverteilung (kategorisiert gemäß den TIMSS-Kompetenzstufen nach Klieme 2000) der FIMS- und TIMSS-Population insgesamt und aufgeschlüsselt nach Gymnasialzweig bzw. Kursform

für die Analysen ausgeschlossen. Diese Schwelle hatte die Konsequenz, dass bei der Skalierung der FIMS-Daten sechs Items (T3-04, T3-14, T5-04, T5-09, T6-13 und T8-14) infolge von Itemtrennschärfen kleiner als 0,15 eliminiert wurden. Somit verblieben 100 Items aus FIMS in den Analysen (EAP-Reliabilität 0,86). Bei der Neuskalierung der TIMSS-Daten wurden vier Items (J02, Trenditem K04, K08 und K15) wegen geringer Itemtrennschärfen entfernt, wodurch 64 Items aus TIMSS in den Analysen (EAP-Reliabilität: 0,82) verblieben. Die Itemstatistiken der Neuskalierungen der FIMS- und TIMSS-Daten finden sich im Online-Supplement.

5.2 Vergleich der Mathematikleistungen in FIMS und TIMSS

Für die TIMSS-Stichprobe ergab sich in unserer Sekundäranalyse ein mittlerer Fähigkeitswert von $M = 455$, $SD = 88$. Diese Werte stimmten mit den Ergebnissen der Primäranalyse von $M = 455$, $SD = 89$ (Baumert et al. 2000c, S. 138) nahezu vollständig überein. Nach dem Mean-Mean-Linking ergab sich für die FIMS-Stichprobe eine mittlere Mathematikleistung von $M = 454$, $SD = 76$. Der Standardfehler

Tab. 2 Mittlere Fähigkeitswerte und deren Verteilungen auf die Kompetenzstufen für TIMSS und FIMS

	TIMSS (Primäranalyse)		TIMSS (Sekundäranalyse)		FIMS	
	Wert	*SE*	Wert	*SE*	Wert	*SE*
M	455	6,9	455	5,2	454	4,4
SD	89	nb	88	5,2	76	3,4
Anteil KS1	24,2 %	nb	24,4 %	2,2 %	22,1 %	2,2 %
Anteil KS2	45,9 %	nb	45,8 %	1,8 %	51,0 %	2,7 %
Anteil KS3	25,3 %	nb	25,3 %	1,7 %	24,5 %	2,2 %
Anteil KS4	4,6 %	nb	4,6 %	0,7 %	2,5 %	0,7 %

KS Kompetenzstufe, *nb* nicht berichtet

Tab. 3 Mittlere Fähigkeitswerte und deren Verteilungen auf die Kompetenzstufen für die Leistungskurse bei TIMSS und die Gymnasien des mathematisch-naturwissenschaftlichen Zweigs bei FIMS

	TIMSS Leistungskurse (Primäranalyse)		TIMSS Leistungskurse (Sekundäranalyse)		FIMS (math.-naturw. Zweig)	
	Wert	*SE*	Wert	*SE*	Wert	*SE*
M	514	nb	521	3,9	497	6,4
SD	71	nb	71	2,5	68	3,5
Anteil KS1	5,7 %	nb	4,4 %	1,0 %	7,3 %	1,9 %
Anteil KS2	37,6 %	nb	33,9 %	3,0 %	44,5 %	4,1 %
Anteil KS3	44,9 %	nb	49,0 %	3,0 %	42,2 %	3,8 %
Anteil KS4	11,7 %	nb	12,7 %	2,0 %	6,0 %	1,7 %

KS Kompetenzstufe, *nb* nicht berichtet

betrug bei FIMS auf der Personenseite $SE = 4{,}4$ und auf der Itemseite $LE = 34{,}9$. Somit ergab sich ein Gesamtfehler von $TE = 35{,}2$. Bei einem Vergleich der Fähigkeitsverteilung auf die Kompetenzstufen nach Klieme (2000) zeigte sich eine sehr ähnliche Verteilung bei FIMS und TIMSS (vgl. Abb. 1 und Tab. 2). Mindestens Kompetenzstufe III, welche gemäß Klieme (2000) als Mindestniveau für die gymnasiale Oberstufe angesehen werden kann, erreichte bei TIMSS 29,9 % der Population, $SE = 1{,}8\,\%$[4], $LE < 0{,}1\,\%$, $TE = 1{,}8\,\%$. Bei FIMS betrug der Anteil der Abiturientinnen und Abiturienten mindestens auf Kompetenzstufe III 26,9 %, $SE = 2{,}5\,\%$, $LE = 14{,}9\,\%$, $TE = 15{,}1\,\%$. Die Differenz der Anteile zwischen FIMS und TIMSS von 3,0 Prozentpunkten fiel nicht statistisch signifikant aus ($SE = 2{,}9\,\%$, $LE = 14{,}9\,\%$, $TE = 15{,}2\,\%$). Selbst unter Ignorierung des Linkingfehlers war der Zugewinn in TIMSS nicht signifikant ($SE = 2{,}9\,\%$, $t = 1{,}02$, $p = 0{,}31$).

Vergleicht man die Mathematikleistungen bei FIMS und TIMSS nach Kursform (Leistungskurs oder Grundkurs) und Zweigen des Gymnasiums (mathematisch-naturwissenschaftlicher Zweig oder sprachlicher Zweig), zeigte sich, dass die durchschnittlichen Mathematikleistungen der Leistungskurse in TIMSS ($M = 521$, $SD = 71$) über den durchschnittlichen Leistungen der Gymnasien des mathematisch-naturwis-

Tab. 4 Mittlere Fähigkeitswerte und deren Verteilungen auf die Kompetenzstufen für die Grundkurse bei TIMSS und die Gymnasien des sprachlichen Zweigs bei FIMS

	TIMSS Grundkurse (Primäranalyse)		TIMSS Grundkurse (Sekundäranalyse)		FIMS (sprachlicher Zweig)	
	Wert	*SE*	Wert	*SE*	Wert	*SE*
M	427	nb	421	6,1	426	5,6
SD	78	nb	77	6,9	68	3,8
Anteil KS1	33,8 %	nb	34,7 %	2,9 %	32,0 %	3,6 %
Anteil KS2	50,3 %	nb	51,9 %	2,5 %	55,2 %	3,5 %
Anteil KS3	15,2 %	nb	13,1 %	1,4 %	12,7 %	2,4 %
Anteil KS4	0,8 %	nb	0,3 %	0,2 %	0,1 %	0,2 %

KS Kompetenzstufe, *nb* nicht berichtet

[4] Die Prozentangaben bei den berichteten Fehlern (*SE, LE* und *TE*) sind im Folgenden immer im Sinne von Prozentpunkten zu interpretieren.

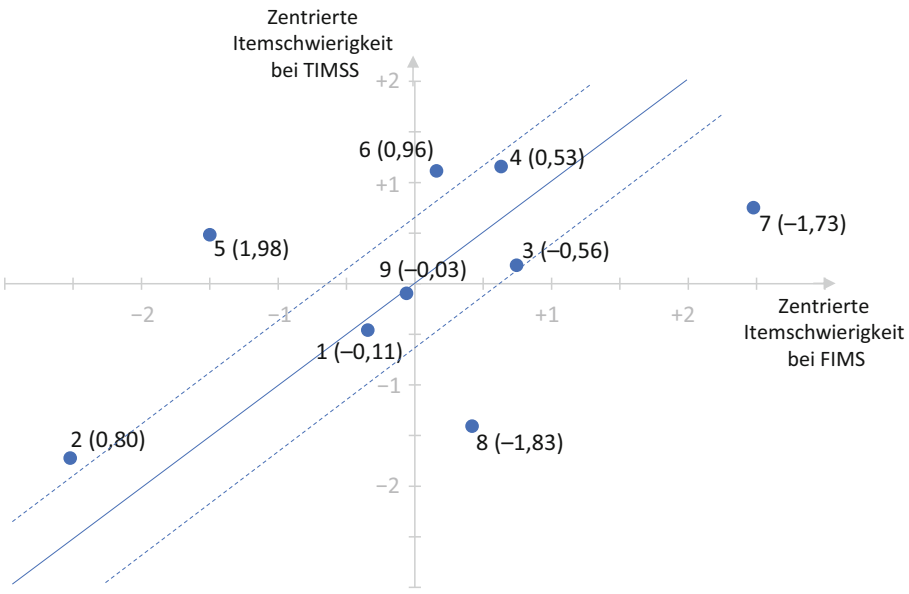

Abb. 2 Graphische Analyse zum differenziellen Itemfunktionieren der Trenditems (Trenditemnummern und der DIF in Klammern)

senschaftlichen Zweigs in FIMS ($M = 497$, $SD = 68$) lagen. Dementsprechend zeigte sich die Verteilung auf die Kompetenzstufen nach Klieme (Abb. 1 und Tab. 3) für die Leistungskurse 1996 günstiger als für die mathematisch-naturwissenschaftlichen Zweige 1964. Dagegen befanden sich die durchschnittlichen Mathematikleistungen der Grundkurse ($M = 421$, $SD = 77$) und der Gymnasien des sprachlichen Zweigs ($M = 426$, $SD = 68$) auf einem ähnlichen Leistungsniveau (vgl. Abb. 1 und Tab. 4).

5.3 Differenzielles Itemfunktionieren der Trenditems

Bei der Analyse der Leistungsergebnisse hatte der Linkingfehler auf Itemseite ein deutlich größeres Ausmaß als der Standardfehler auf der Personenseite. Um dieses Resultat näher zu untersuchen, wurden die in der Trendanalyse verbliebenen neun Trenditems (vgl. Anhang) auf differenzielles Itemfunktionieren (DIF; Holland und Wainer 1993) untersucht. Diese DIF-Analysen dienen nicht nur dazu, um aus psychometrischer Sicht die Modellpassung zu evaluieren, sondern sollen auch einen Eindruck vermitteln, für welche mathematischen Teilbereiche differenzielle Trendeffekte auftraten. Hierbei zeigte sich (vgl. Abb. 2 und Items im Anhang), dass die Trenditems 1 (T3-17 bzw. L01), 3 (T5-16 bzw. L02), 4 (T6-15 bzw. L05) und 9 (T8-03 bzw. K11) einen DIF unter 0,6 Logit aufwiesen, was ungefähr der C-DIF-Klassifikation des Educational Testing Service entspricht (Zieky 1993). Das Trenditems 2 (T5-06 bzw. K01) und das Trenditem 6 (T9-10 bzw. L07) hatten einen DIF zwischen 0,6 und 1,0 Logit. Die drei weiteren Trenditems 5 (T9-03 bzw. I06), 7 (T6-10 bzw. L03) und 8 (T9-13 bzw. J08) hatten substanzielle DIF-Werte über 1,0 Logit.

Eine qualitative Analyse der drei Items mit DIF-Werten über einen Logit (Trenditems 5, 7 und 8) zeigte, dass der DIF möglicherweise durch die unterschiedliche unterrichtliche Nähe der Items zu den jeweiligen Erhebungszeitpunkten erklärt werden kann. So beschäftigt sich das Trenditem 9 mit der Subtraktion zweier Vektoren. Im Jahre 1996 stellte die Analytische Geometrie in vektorieller Form ein Kernelement des Oberstufenunterrichts dar, wie damals eingesetzte Schulbücher (Schmid und Schweizer 1986) zeigen. Dagegen war die Analytische Geometrie in vektorieller Form 1964 nicht in gleichem Maße etabliert. Wie Analysen von damals eingesetzten Schulbüchern (z. B. Schweizer und Lambacher 1959; Wolff 1959) zeigen, dominierte in dieser Zeit im Bereich der Analytischen Geometrie die Untersuchung von Kegelschnitten (Kreis, Ellipse, Hyperbel, Parabel), welche in den im Jahr 1996 verwendeten Schulbüchern kaum noch Niederschlag fanden. Zusammengefasst kann daher festgestellt werden, dass das Trenditem 8 im Jahre 1996 eine deutlich höhere unterrichtliche Nähe aufwies als im Jahre 1964.

Ähnliche Schlussfolgerungen, wenn auch nicht in dieser Eindeutigkeit, können für das Trenditem 8 gezogen werden. Inhalt dieses Items ist die Anwendung der Exponentialfunktion in einem Sachkontext. Wie eine Schulbuchanalyse der klassischen gymnasialen Schulbuchreihe (*Lambacher-Schweizer*) zeigt, wurde die Exponential- und Logarithmusfunktion im *Lambacher-Schweizer* für den Abiturjahrgang 1964 (Schweizer und Lambacher 1959) im vorletzten von 41 Kapiteln auf lediglich drei von 133 Seiten (2,2 %) behandelt. Dagegen wird dieses Thema im *Lambacher-Schweizer* für den Abiturjahrgang 1996 (Schmid und Schweizer 1990) zwar auch gegen Ende des Buches, aber mit einer größeren Tiefe behandelt. Hier nimmt es 25 Seiten von insgesamt 243 Seiten (10,3 %) ein. Trotz aller Unzulänglichkeiten dieser quantitativen Schulbuchanalyse deuten diese unterschiedlichen Anteile darauf hin, dass das Thema Exponential- und Logarithmusfunktion 1996 intensiver behandelt wurde als 1964.

Mit umgekehrtem Vorzeichen deuten die Analyse des Schulbuches Lambacher-Schweizer darauf hin, dass das Trenditem 5 im Jahr 1964 eine größere unterrichtliche Nähe aufwies als im Jahr 1996. Zur Lösung des Trenditems 5 muss eine Wurzelfunktion mit Hilfe der Kettenregel differenziert werden. Im Lambacher-Schweizer für 1964 (Schweizer und Lambacher 1959) wird die Kettenregel und die Ableitung der Wurzelfunktion direkt nach der Einführung in die Differentialrechnung in drei von 24 Kapiteln behandelt, sodass insgesamt auf sieben von 133 Seiten (5,2 %) die für das Trenditem 5 relevanten mathematischen Verfahren behandelt wurden. Dagegen werden im *Lambacher-Schweizer* für 1996 (Schmid und Schweizer 1990) die Kettenregel und die Ableitung der Wurzelfunktion erst in der zweiten Hälfte des Buches in einem Kapitel zur „Weiterführung der Differentialrechnung“ auf neun von 243 Seiten (3,7 %) thematisiert. Auch wenn sich die Sachlage bezüglich des Trenditems 5 nicht so eindeutig wie für die Trenditems 7 und 8 darstellt, so lässt sich anhand der dargestellten Indizien vermuten, dass 1964 die Ableitung der Wurzelfunktion mit Hilfe der Kettenregel einen größeren Raum im Mathematikunterricht einnahm als im Jahr 1996.

6 Diskussion

6.1 Zusammenfassung der Ergebnisse

Die Analysen zeigen, dass sich insgesamt kein negativer Trend in den Mathematikleistungen der Abiturientinnen und Abiturienten von 1964 bis 1996 finden ließ. Sowohl die mittleren Testleistungen (FIMS: 454; TIMSS: 455) als auch der Anteil der Population, der auf dem TIMSS-Kompetenzstufenmodell nach Klieme (2000) mindestens Stufe III erreichte (FIMS: 26,9 %, TIMSS: 29,9 %), unterschieden sich nicht signifikant. Deskriptiv war allerdings festzustellen, dass die Streuung der Mathematikleistungen in der TIMSS-Population größer war als in der FIMS-Population. Bei der vergleichenden Betrachtung der Schulzweige bzw. der Kursformen zeigte sich, dass die Gymnasien des mathematisch-naturwissenschaftlichen Zweigs 1964 geringere Mathematikleistungen zeigten als die Leistungskurse 1996. Dagegen befand sich das Leistungsniveau der Gymnasien der sprachlichen Zweige 1964 auf einem vergleichbaren Niveau wie das der Grundkurse 1996. Aus bildungshistorischer Perspektive liefern die Analysen Indizien dafür, dass für das Fach Mathematik die Neugestaltung der gymnasialen Oberstufe in den 70iger Jahren möglicherweise keine substanzielle Zäsur darstellte. Stattdessen zeigte sich für das Fach Mathematik eine deutliche Kontinuität der Rahmenbedingungen (z. B. Stundenzahl) und der Leistungen von Schulzweigen (mathematisch-naturwissenschaftlich vs. sprachlich) zu den Kursformen (Leistungskurs vs. Grundkurs).

Die Trendschätzung zeigte ein substanzielles Ausmaß an Unsicherheit. Der aus der Personenauswahl entstehende Standardfehler *SE* für den Anteil der Abiturientinnen und Abiturienten bei FIMS, die das Mindestniveau erreichten, betrug zwar moderate 2,5 %. Dagegen überstieg der Standardfehler *LE*, der durch die Auswahl der Trenditems verursacht wurde, den Standardfehler auf Grund der Personenauswahl deutlich. Das bedeutet, dass die Unsicherheit in den Trendschätzungen vor allem durch die begrenzte Anzahl von nur neun Trenditems und weniger durch die Stichprobengröße verursacht wurde.

Die DIF-Analysen zeigten, dass es zum Teil substanzielle itemspezifische Unterschiede in der Trendentwicklung gab. So wiesen fünf Items einen substanziellen DIF (d. h. DIF $>0{,}6$ Logit) auf. Bei einer qualitativen Analyse von DIF-Items stellte sich heraus, dass die unterschiedliche unterrichtliche Nähe der DIF-Items einen Erklärungsansatz darstellen könnte.

6.2 Zusammenhang Bildungsbeteiligung und Mathematikleistungen

Die Frage nach dem Zusammenhang zwischen der Bildungsbeteiligung und den Mathematikleistungen wurde nicht explizit als Forschungsfrage formuliert, da die vorhandenen Daten keine kausalen Aussagen erlauben. Trotzdem erscheint eine Reflexion der Analyseergebnisse vor dem Hintergrund dieser Frage berechtigt. Rein deskriptiv betrachtet ist ein Zusammenhang zwischen Bildungsbeteiligung und Mathematikleistungen von Abiturientinnen und Abiturienten zunächst nicht feststellbar. Zwar verdreifachte sich die Abiturientenquote im Zeitraum von 1964 bis 1996, ein entsprechendes Absinken der Mathematikleistungen, wie der Hamburger Landschul-

rat Ernst Matthewes im Jahre 1964 prognostiziert (vgl. Theoretischer Hintergrund), ist aber nicht feststellbar.

Zur Erklärung dieses empirischen Befundes sind zwei Ansätze denkbar. Zum einen könnte die These formuliert werden, dass sehr wohl ein negativer kausaler Zusammenhang zwischen der höheren Bildungsbeteiligung und den Mathematikleistungen besteht, sich dieser Effekt in der Leistungsentwicklung von 1964 bis 1996 aber nicht zeigt, da andere kompensatorische Effekte den negativen Effekt der Bildungsexpansion ausgeglichen haben. So wäre eine plausible Vermutung, dass sich die Lehr- und Lernbedingungen von 1964 zu 1996 verbessert haben (z. B. besserer Mathematikunterricht und bessere sozioökonomische Bedingungen im Jahr 1996 als im Jahr 1964). Somit wäre der negative kausale Zusammenhang wegen kompensatorischer, konfundierender Effekte in den Ergebnissen nicht sichtbar geworden.

Zum anderen könnte die These aufgestellt werden, dass tatsächlich kein bzw. kaum ein kausaler Zusammenhang zwischen der Bildungsbeteiligung und den Mathematikleistungen in der Sekundarstufe II besteht. Die Annahme, dass sich eine Bildungsexpansion negativ auf das Leistungsniveau auswirkt, beruht zu einem großen Teil auf der Annahme, dass sich das durchschnittliche Intelligenzniveau der Schülerschaft bei einer höheren Bildungsbeteiligung verringert. Wie allerdings Daten aus der Sekundarstufe I andeuten, ist diese Annahme nicht unbedingt berechtigt. So nahm der IQ-Mittelwert in der Gymnasialpopulation der 7. Klassen in NRW von 1968 ($M = 108{,}9$) bis 1991 ($M = 112{,}9$) um 4,0 IQ-Punkte zu (Becker et al. 2006), obwohl in diesem Zeitraum das Gymnasium einen deutlichen Zuwachs der Schülerschaft zu verzeichnen hatte. Becker et al. (2006) vermuten, dass dieses Ergebnis dem sogenannten Flynn-Effekt zuzuschreiben ist, der die Tatsache beschreibt, dass in den Industrieländern ein Intelligenzzuwachs über die Generationen gemessen wurde (Flynn 1987). Neben dem Flynn-Effekt könnte eine weitere Erklärung sein, dass in den 1960er-Jahren nicht unbedingt nur die Schülerinnen und Schüler mit einem hohen kognitiven Potenzial (d. h. hohen IQ-Werten) auf das Gymnasium wechselten, sondern die Übergangsentscheidung am Ende der Grundschule maßgeblich von anderen Faktoren (z. B. Geschlecht, Erwartungen des Elternhauses, sozio-ökonomische Bedingungen) beeinflusst wurde.

Außerdem zeigen empirische Studien, dass die Intelligenz zwar ein wichtiger Prädiktor für Schülerleistungen ist, der aber durch andere Faktoren (z. B. Vorwissen) kompensiert werden kann. Dieser Argumentation folgend sind andere Faktoren, wie zum Beispiel die Stundenzahl und das Vorwissen, bedeutendere Faktoren für die Leistungsentwicklung als das Ausmaß der Bildungsbeteiligung. Empirische Evidenz für diese These des Nichtzusammenhangs könnten die Ergebnisse von Slowenien aus TIMSS 1995 bieten. In Slowenien erhielten 75 % des einschlägigen Altersjahrgangs Mathematikunterricht in voruniversitärer Mathematik (Baumert et al. 2000c). Trotz dieses im Vergleich zu Deutschland (25 %) hohen Anteils befanden sich die durchschnittlichen Leistungen von Slowenien mit 475 Punkten sogar noch über den durchschnittlichen Leistungen in Deutschland (455 Punkte). Somit zeigten die Ergebnisse Sloweniens nach Baumert et al. (2000c), dass „eine Öffnung der Wege zur Hochschulreife [...] keineswegs zur Leistungsnivellierung oder Senkung des Niveaus führen [muss]“ (S. 146).

6.3 Limitationen

FIMS wurde nicht in der gesamten Bundesrepublik, sondern nur in den beiden Bundesländern Hessen und Schleswig-Holstein durchgeführt. Eine Beschränkung der TIMSS-Analysen nur auf diese beiden Bundesländer wäre aus statistischer Sicht problematisch, da beim Sampling in TIMSS nicht nach Bundesländern stratifiziert wurde (Baumert et al. 2000a). Neben einer deutlichen Verkleinerung der Stichprobe hätte eine Beschränkung auf die beiden Bundesländer zu einer nicht repräsentativen Stichprobe geführt. Außerdem zeigen Ergebnisse für 15-jährige Schülerinnen und Schülern aus der PISA-Erweiterungsstudie von 2000 (Neubrand und Klieme 2002) und 2003 (Neubrand et al. 2005), dass sich die Mathematikleistungen in Hessen (2000: 486; 2003: 497) und Schleswig-Holstein (2000: 490; 2003: 497) jeweils nicht signifikant vom Bundesdurchschnitt (2000: 490; 2003: 503) unterschieden. Unter der Annahme, dass auch für die Sekundarstufe II die Mathematikleistungen in den beiden Bundesländer Schleswig-Holstein und Hessen nahe am bundesdeutschen Mittel liegen, können die bundesweiten TIMSS-Ergebnisse von 1996 als Proxy für die Mathematikleistungen von Abiturientinnen und Abiturienten in den beiden Bundesländern Schleswig-Holstein und Hessen verwendet werden. Gleichwohl ist nicht vollständig auszuschließen, dass diese Annahme nicht zutreffen ist und sich die Kompetenzentwicklung in den beiden Bundesländern für die Sekundarstufe I und die Sekundarstufe II um die Jahrtausendwende unterschieden hat, sodass diese Limitation eine zusätzliche Unsicherheit für die Trendergebnisse der beiden Bundesländer Hessen und Schleswig-Holstein darstellt.

Für das Linking der FIMS und TIMSS-Ergebnisse standen nur neun Trenditems zur Verfügung. Bei den Trenditems traten zum Teil deutliche DIF-Effekte auf, woraus sich im Allgemeinen aber nicht ableiten lässt, welche Items den Trend „real“ beschreiben. Daher wurden keine Items aus dem Linking entfernt, weil unklar ist, weshalb eine Entfernung von Items zu einer höheren Validität der Trendschätzung führen sollte. Um aber eine Abschätzung über die Unsicherheit der Trendschätzung zu erhalten, wurde Analysen auf der Basis eines Itemsamplings durchgeführt. Hier zeigte sich, dass der Linkfehler, der durch die Itemauswahl entstand, durchaus bedeutend ist und zu einem nicht unerheblichen Teil von der Itemauswahl abhängig ist. Allerdings kann in unserer Studie auch keine empirische Evidenz für Leistungsunterschiede 1964 und 1996 gefunden werden, was angesichts der stark verbreiteten Annahmen über den negativen Zusammenhang von Bildungsexpansion und Mathematikleistungen und einer allgemein negativen Trendentwicklung der Mathematikleistungen bereits bemerkenswert ist. Werden die neun Trenditems als repräsentative Zufallsauswahl aus der Menge möglicher Items betrachtet, so besitzt wegen der Erwartungstreue des Mittelwerts die Schlussfolgerung, dass sich die Mathematikleistungen in FIMS und TIMSS nicht substanziell unterschieden, trotz aller Unsicherheit die höchste Plausibilität.

Eine weitere Limitation stellt die leichte Modifizierung der Itemformulierungen der Trenditems von FIMS nach TIMSS dar. So waren sieben der neun Items in der englischen Fassung bei FIMS und TIMSS sprachlich vollständig identisch, während zwei Items in der englischen Originalfassung sehr leichte sprachliche Modifikationen aufwiesen. Außerdem wurden bei der Übersetzung der Items aus der englischen

Originalfassung ins Deutsche von FIMS zu TIMSS zum Teil leichte sprachliche Übersetzungsanpassungen vorgenommen. Diese jeweiligen sprachlichen Anpassungen, die den mathematischen Inhalt und auch die Antwortmöglichkeiten vollständig unberührt ließen, könnten jedoch Auswirkungen auf die Itemschwierigkeiten gehabt haben. Allerdings wiesen sie ein so geringes Ausmaß auf, dass ein substanzieller Einfluss auf die Itemschwierigkeiten nicht angenommen wird.

Schließlich stellt sich die Frage, ob der vorliegende Leistungsvergleich als *fair* oder *angemessen* betrachtet werden kann, da zwischen den beiden Erhebungszeitpunkten 32 Jahre liegen und die Zusammensetzung und der Hintergrund der Testpersonen sich deutlich unterschied. Es ist fraglich, ob ein Leistungsvergleich zwischen Generation jemals als *fair* betrachtet werden kann. Beispielsweise wuchsen die Abiturientinnen und Abiturienten von 1996 im Vergleich zu ihren Altersgenossen von 1964 im Allgemeinen in gesicherteren ökonomischen Verhältnissen auf und auch war die Ausstattung des Bildungssystems 1996 besser als 1964 (vgl. Picht 1964). Auf der anderen Seite gab es für die Abiturientinnen und Abiturienten von 1996 möglicherweise nachteilige Bedingungsfaktoren (z. B. erhöhter Medienkonsum) oder zusätzlichen Anforderungen (z. B. zusätzliche Beschäftigung mit dem Themengebiet Stochastik), denen die Abiturientengeneration von 1964 nicht ausgesetzt waren. Daher ist es grundsätzlich schwierig zu beurteilen, ob ein Trendvergleich als *fair* betrachtet werden kann. Würde man Fairness als notwendiges Kriterium für einen Leistungsvergleich voraussetzen, so wären auch internationale Leistungsvergleiche wie beispielsweise bei PISA zwischen Ländern mit sehr unterschiedlichen sozioökonomischen Rahmenbedingungen wie Deutschland und Rumänien abzulehnen. Daher können die vorliegenden Analysen zur Fairness des Leistungsvergleichs keine Aussage treffen, gleichwohl können sie *angemessen* sein, da sie Entwicklungen deskriptiv aufzeigen und möglicherweise Fehlwahrnehmungen (z. B. potenziell nostalgisch verzerrte Wahrnehmung der Leistungen früherer Abiturientengenerationen) aufdecken. Um allerdings nicht lediglich auf der deskriptiven Ebene zu verbleiben, wurden die Trendanalysen durch inhaltliche Analysen der Trenditems, damals verwendeter Schulbücher und geltender Curricula ergänzt, um differenzielle Effekte trotz aller Limitationen auch dieser Analysen zumindest im Ansatz erklären zu können.

6.4 Forschungsdesiderate

Die empirische Datenlage zu den Mathematikleistungen in der gymnasialen Oberstufe und deren Trendentwicklung ist unzureichend, da kein Bildungsmonitoring für die Sekundarstufe II stattfindet. Daher können über das aktuelle Leistungsniveau der Schülerinnen und Schüler in der gymnasialen Oberstufe keine gesicherten Aussagen gemacht werden. Dieses ist umso überraschender, als dass die Einrichtung eines einheitlichen IQB-Aufgabenpools für die Abiturprüfungen bei einer geeigneten Erfassung der Abiturergebnisse eine Evaluierung der Mathematikleistungen ermöglichen würde. Bisher werden aber die Ergebnisse der Abiturprüfungen nicht detailliert erfasst, sondern in den Bundesländern werden lediglich Durchschnittsnoten veröffentlicht. Daher gäbe es die Möglichkeit, mit Hilfe der jährlichen Abiturprüfungen die Mathematikleistungen der Abiturientinnen und Abiturienten genauer zu

evaluieren. Kahnert (2014) konnte zeigen, dass der TIMSS-Test zur voruniversitären Mathematik und die Mathematikaufgaben aus dem Zentralabitur 2011 in Nordrhein-Westfalen ein eindimensionales Konstrukt darstellten. Dies bedeutet, dass Abituraufgaben vergleichbare Anforderungen wie die TIMSS-Items zur voruniversitären Mathematik stellen und somit die Zentralabituraufgaben für ein Bildungsmonitoring psychometrisch durchaus geeignet wären.

In der vorliegenden Trendanalyse stellte sich die limitierte Anzahl an Linkitems als größter Faktor der Unsicherheit in der Trendschätzung heraus. Daher könnte eine Linkingstudie, in der Items aus FIMS und TIMSS gemeinsam eingesetzt werden, eine breitere empirische Grundlage für das Linking der beiden Erhebungen erzeugen. Würde diese Linkingstudie mit einer repräsentativen Stichprobe aus Hessen und Schleswig-Holstein durchgeführt werden, könnten Aussagen über die Trendentwicklung der Mathematikleistungen von Abiturientinnen und Abiturienten in Hessen und Schleswig-Holstein von 1964 bis in die Gegenwart gemacht werden.

Ein weiteres, in der Bildungsforschung bisher nicht untersuchtes Feld ist die Wahrnehmung der Vergangenheit durch im Bildungsbereich tätige Personen. Wie in der Einleitung angedeutet, könnte beispielsweise Nostalgie eine mögliche Ursache für potenziell verzerrte Wahrnehmungen der Mathematikleistungen vergangener Generationen sein. Eine andere mögliche Erklärung wäre, dass der Expertisezuwachs von Lehrkräften und Hochschullehrenden während der Berufslaufbahn die Wahrnehmung der Mathematikleistungen der Schülerinnen und Schüler bzw. Studierenden negativ beeinflusst. Systematischere Forschung für den Bildungsbereich könnte möglicherweise auf Forschungsergebnissen aus der *historischen Memorik* (Fried 2004) oder der *interdisziplinären Gedächtnisforschung* (*memory studies*, Dutceac Segesten und Wüstenberg 2017) aufbauen und somit Wirkmechanismen von Erinnerungsprozessen im Bildungsbereich identifizieren.

6.5 Praktische Implikationen

Die Ergebnisse der vorliegenden Trendanalyse legen nahe, dass auch in den 1960er-Jahren ein Großteil der Abiturientinnen und Abiturienten nicht das kriteriale Mindestniveau gemäß dem Kompetenzstufenmodell nach Klieme (2000) erreichte. Dies bedeutet, dass entweder das kriteriale Mindestniveau zu hoch angesetzt wurde oder dass der Mathematikunterricht in der gymnasialen Oberstufe seit Jahrzehnten seine selbstgesteckten Ziele nicht erreicht (vgl. Rolfes et al. 2021). Trotz der Bestrebungen zur Jahrtausendwende, im Mathematikunterricht im Zuge der Kompetenzorientierung eine verstärkt verständnisorientierte statt kalkülorientierte Arbeitsweise zu etablieren, neigt der Mathematikunterricht in der gymnasialen Oberstufe nach Ansicht der Autoren immer noch zu einem gewissen Enzyklopädismus der Begriffe und Verfahren. Durch die verstärkte Einbeziehung der Stochastik neben der Analysis und der Analytischen Geometrie/Linearen Algebra in den Oberstufenunterricht hat sich dieser Umstand in den letzten zwei Jahrzehnten sogar noch erhöht, da von den Schülerinnen und Schülern am Ende der gymnasialen Oberstufe in einem durchaus umfangreichen Maße eine Beherrschung von Begriffen und Verfahren aus diesen drei Inhaltsbereichen erwartet wird. Eine kritische Reflexion der erwarteten curricularen Mathematikkompetenzen im Lichte der empirischen Ergebnisse könn-

te dazu beitragen, die essenziellen Bildungsziele des Mathematikunterrichts in der gymnasialen Oberstufe herauszuarbeiten (vgl. Rolfes et al. 2022) und realistischere Kompetenzerwartungen an den Mathematikunterricht und die Abiturientinnen und Abiturienten zu formulieren.

Zusatzmaterial online Zusätzliche Informationen sind in der Online-Version dieses Artikels (https://doi.org/10.1007/s11618-023-01176-6) enthalten.

Funding Open Access funding enabled and organized by Projekt DEAL.

Interessenkonflikt T. Rolfes, A. Robitzsch und A. Heinze geben an, dass kein Interessenkonflikt besteht.

Literatur

(1964). Die letzte Hürde. *Der Spiegel*, *18*(50), 73–87.

(2017). Mathematikunterricht und Kompetenzorientierung – ein offener Brief. https://www.tagesspiegel.de/wissen/downloads/offener-brief-der-mathematiker. Zugegriffen: 15. Juli 2023.

Batcho, K. I. (1995). Nostalgia: a psychological perspective. *Perceptual and Motor Skills*, *80*(1), 131–143.

Baumert, J., & Köller, O. (2000). Motivation, Fachwahlen, selbstreguliertes Lernen und Fachleistungen im Mathematik- und Physikunterricht der gymnasialen Oberstufe. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe. TIMSS/III Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (Bd. 2, S. 181–213). Opladen: Leske + Budrich.

Baumert, J., & Lehmann, R. (1997). *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich*. Opladen: Leske + Budrich. https://doi.org/10.1007/978-3-322-95096-3.

Baumert, J., Bos, W., Klieme, E., Lehmann, R., Lehrke, M., Hosenfeld, I., Neubrand, J., & Watermann, R. (1999). *Testaufgaben zu TIMSS/III. Mathematisch-naturwissenschaftliche Grundbildung und voruniversitäre Mathematik und Physik der Abschlußklassen der Sekundarstufe II (Population 3)*. Berlin: Max-Planck-Institut für Bildungsforschung.

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

Baumert, J., Bos, W., & Lehmann, R. (Hrsg.). (2000b). *Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe. TIMSS/III Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (Bd. 2). Opladen: Leske + Budrich.

Baumert, J., Bos, W., & Watermann, R. (2000c). Fachleistungen im voruniversitären Mathematik- und Physikunterricht im internationalen Vergleich. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe. TIMSS/III Dritte*

Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn (Bd. 2, S. 129–180). Opladen: Leske + Budrich.
Becker, M., Trautwein, U., Lüdtke, O., Cortina, K. S., & Baumert, J. (2006). Bildungsexpansion und kognitive Mobilisierung. In A. Hadjar & R. Becker (Hrsg.), *Die Bildungsexpansion* (S. 63–89). Wiesbaden: VS.
Behnke, H. (1965). Die Pflichten der Universität gegenüber dem Gymnasium. *Mathematisch-physikalische Semesterberichte*, *11*, 1–19.
Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.
Büchter, A., & Henn, H.-W. (2015). Schulmathematik und Realität – Verstehen durch Anwenden. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme & H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (S. 19–49). Berlin: Springer.
Buschhüter, D., Spoden, C., & Borowski, A. (2016). Mathematische Kenntnisse und Fähigkeiten von Physikstudierenden zu Studienbeginn. *Zeitschrift für Didaktik der Naturwissenschaften*, *22*(1), 61–75. https://doi.org/10.1007/s40573-016-0041-4.
DMV (Deutsche Mathematiker-Vereinigung), DPG, GDCh, MNU & VDB (1982). Rettet die mathematisch-naturwissenschaftliche Bildung. *Physikalische Blätter*, *38*(1), 25.
Dutceac Segesten, A., & Wüstenberg, J. (2017). Memory studies. The state of an emergent field. *Memory Studies*, *10*(4), 474–489. https://doi.org/10.1177/1750698016655394.
Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*(2), 171–191.
Fried, J. (2004). *Der Schleier der Erinnerung. Grundzüge einer historischen Memorik*. München: Beck.
Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement*, *8*(3), 305–322.
Hessisches Kultusministerium (1991). *Kursstrukturpläne: gymnasiale Oberstufe. Aufgabenfeld 3: 1. Mathematik*. Wiesbaden: Hessisches Kultusministerium.
HMEV (Hessisches Ministerium für Erziehung und Volksbildung) (1957). Bildungspläne für die allgemeinbildenden Schulen im Lande Hessen. Mathematik. *Amtsblatt des hessischen Ministers für Erziehung und Volksbildung*, *10*(4), 546–554.
HMEV (1956). Bildungspläne für die allgemeinbildenden Schulen im Lande Hessen. Der Bildungsplan des Gymnasiums. *Amtsblatt des hessischen Ministers für Erziehung und Volksbildung*, *9*, 104–117.
Holland, P. W., & Wainer, H. (Hrsg.). (1993). *Differential item functioning*. Hillsdale: Erlbaum.
Husén, T. (Hrsg.). (1967). *International study of achievement in mathematics. A comparison of twelve countries*. Stockholm: Almqvist & Wiksell.
Kahnert, J. (2014). *Das Zentralabitur im Fach Mathematik. Eine empirische Analyse von Abitur- und TIMSS-Daten im Vergleich*. Münster: Waxmann.
Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe. TIMSS/III Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (Bd. 2, S. 57–128). Opladen: Leske + Budrich.
KMK (1989). *Einheitliche Prüfungsanforderungen in der Abiturprüfung Mathematik. (Beschluss der Kultusministerkonferenz vom 01.12.1989).* Neuwied: Luchterhand.
KMK (1995). *Weiterentwicklung der Prinzipien der gymnasialen Oberstufe und des Abiturs. Abschlussbericht der von der Kultusministerkonferenz eingesetzten Expertenkommission*. Kiel: Schmidt & Klaunig.
KMK (2019). *Schüler, Klassen, Lehrer und Absolventen der Schulen 2008 bis 2017*. Berlin: KMK.
KMSH (Kultusminister des Landes Schleswig-Holstein) (1955). *Lehrplanrichtlinien für die Gymnasien*. Lübeck: Antäus.
KMSH (1956). Stundentafeln für die Gymnasien in Schleswig-Holstein. *Nachrichtenblatt für das schleswig-holsteinische Schulwesen*, *1*, 1–4.
KMSH (1980). Stundentafeln der allgemeinbildenden Gymnasien. *Nachrichtenblatt des Kultusministers des Landes Schleswig-Holstein*, *11*, 202–210.
KMSH (1986). *Gymnasium. Übersichten zu den Lehrplänen*. Kiel: Kultusministerium des Landes Schleswig-Holstein.
Köhler, H., Lundgreen, P., Rochow, T., & Schallmann, J. (2014). *Allgemein bildende Schulen in der Bundesrepublik Deutschland 1949–2010* (Datenhandbuch zur deutschen Bildungsgeschichte, Bd. 7). Göttingen: Vandenhoeck & Ruprecht.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3. Aufl.). New York: Springer. https://doi.org/10.1007/978-1-4939-0317-7.

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *The Stata Journal*, *10*(2), 165–199.

Lambacher, T., & Schweizer, W. (Hrsg.). (1959). *Analysis* (Kurzausgabe). Stuttgart: Klett.

Land Hessen (1982). *Gesetz über die gymnasiale Oberstufe und zur Änderung anderer Vorschriften. Gesetz- und Verordnungsblatt für das Land Hessen – Teil I, 10* (S. 140–143).

Leboe, J. P., & Ansons, T. L. (2006). On misattributing good remembering to a happy past: An investigation into the cognitive roots of nostalgia. *Emotion*, *6*(4), 596–610. https://doi.org/10.1037/1528-3542.6.4.596.

Lenné, H. (1975). *Analyse der Mathematikdidaktik in Deutschland* (2. Aufl.). Stuttgart: Klett.

Leutner, D., Hartig, J., & Jude, N. (2008). Measuring competencies: Introduction to concepts and questions of assessment in education. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 177–192). Göttingen: Hogrefe & Huber.

Majoros, E., Rosén, M., Johansson, S., & Gustafsson, J.-E. (2021). Measures of long-term trends in mathematics: linking large-scale assessments over 50 years. *Educational Assessment, Evaluation and Accountability*, *33*(1), 71–103. https://doi.org/10.1007/s11092-021-09353-z.

Martin, M. O., & Kelly, D. L. (Hrsg.). (1998). *mplementation and analysis. Final year of secondary school (Population 3). Third international mathematics and science study* (Technical report, Bd. III). Chestnut Hill: Boston College.

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, *8*(3), 323–335.

Mullis, I. V., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS advanced 2015 international results in advanced mathematics and physics*. Chestnut Hill: Boston College.

National Center for Education Statistics (2012). *The nation's report card: trends in academic progress 2012 (NCES 2013 456)*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Neubrand, M., & Klieme, E. (2002). Mathematische Grundbildung. In J. Baumert & M. Neubrand (Hrsg.), *PISA 2000 – die Länder der Bundesrepublik Deutschland im Vergleich* (S. 95–127). Opladen: Leske + Budrich.

Neubrand, M., Blum, W., Ehmke, T., Jordan, A., Senkbeil, M., Ulfig, F., & Carstensen, C. H. (2005). Mathematische Kompetenz im Ländervergleich. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, J. Rost & U. Schiefele (Hrsg.), *PISA 2003. Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche?* (S. 51–84). Münster: Waxmann.

Neumann, M., & Trautwein, U. (2019). Sekundarbereich II und der Erwerb der Hochschulzugangsberechtigung. In O. Köller, M. Hasselhorn, F. W. Hesse, K. Maaz, J. Schrader & H. Solga (Hrsg.), *Das Bildungswesen in Deutschland. Bestand und Potenziale* (S. 533–564). Bad Heilbrunn: Klinkhardt.

Picht, G. (1964). *Die deutsche Bildungskatastrophe*. Olten: Walter.

Pinker, S. (2018). *Enlightenment now: The case for reason, science, humanism, and progress*. New York: Viking.

Postlethwaite, T. N. (Hrsg.). (1968). *IEA Leistungsmessung in der Schule. Eine internationale Untersuchung am Beispiel des Mathematikunterrichts*. Frankfurt a. M.: Diesterweg.

Robitaille, D. F., & Taylor, A. R. (1989). Changes in patterns of achievement between the first and second mathematics studies. In D. F. Robitaille & R. A. Garden (Hrsg.), *The IEA study of mathematics II: Contexts and outcomes of school mathematics* (S. 153–177). Oxford: Pergamon.

Robitzsch, A. (2021). Robust and nonrobust linking of two groups for the Rasch model with balanced and unbalanced random DIF: a comparative simulation study and the simultaneous assessment of standard errors and linking errors with resampling techniques. *Symmetry*, *13*(11), 2198. https://doi.org/10.3390/sym13112198.

Robitzsch, A., & Lüdtke, O. (2019). Linking errors in international large-scale assessments: calculation of standard errors for trend estimation. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 444–465. https://doi.org/10.1080/0969594X.2018.1433633.

Robitzsch, A., Dörfler, T., Pfost, M., & Artelt, C. (2011). Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *43*(4), 213–227. https://doi.org/10.1026/0049-8637/a000052.

Rolfes, T., Lindmeier, A., & Heinze, A. (2021). Mathematikleistungen von Schülerinnen und Schülern der gymnasialen Oberstufe in Deutschland: Ein Review und eine Sekundäranalyse der Schulleistungsstudien seit 1995. *Journal für Mathematik-Didaktik*, *42*(2), 395–429. https://doi.org/10.1007/s13138-020-00180-1.

Rolfes, T., Rach, S., Ufer, S., & Heinze, A. (Hrsg.). (2022). *Das Fach Mathematik in der gymnasialen Oberstufe*. Münster: Waxmann. https://doi.org/10.31244/9783830996019.
Rosling, H., Rosling, O., & Rönnlund, A.R. (2018). *Factfulness: ten reasons we're wrong about the world and why things are better than you think*. London: Sceptre.
Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Huber.
Rupp, A.A., & Zumbo, B.D. (2006). Understanding Parameter Invariance in Unidimensional IRT Models. *Educational and Psychological Measurement*, *66*(1), 63–84. https://doi.org/10.1177/0013164404273942.
Schmid, A., & Schweizer, W. (Hrsg.). (1986). *LS Mathematik. Analytische Geometrie mit Linearer Algebra. Grundkurs*. Stuttgart: Klett.
Schmid, A., & Schweizer, W. (Hrsg.). (1990). *LS Mathematik. Analysis. Grundkurs. Gesamtausgabe*. Stuttgart: Klett.
Schultze, W. (Hrsg.). (1968). *Schulen in Europa. Band I: Teil A*. Weinheim: Beltz.
Schultze, W., & Riemenschneider, L. (1967). *Eine vergleichende Studie über die Ergebnisse des Mathematikunterrichts in zwölf Ländern*. Frankfurt a. M.: Deutsches Institut für Internationale Pädagogische Forschung.
Schweizer, W., & Lambacher, T. (Hrsg.). (1959). *Analytische Geometrie (Kurzausgabe)* (4. Aufl.). Stuttgart: Klett.
Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (1972). Beschlüsse der Kultusministerkonferenz. Vereinbarungen zur Neugestaltung der gymnasialen Oberstufe in der Sekundarstufe II. https://www.kmk.org/fileadmin/Dateien/pdf/Bildung/AllgBildung/176_Vereinb_Gestalt_Gym_Ob_Sek_II-1972_01.pdf. Zugegriffen: 15.07.2023.
Stanat, P., Becker-Mrotzek, M., Blum, W., & Tesch, B. (2016). Vergleichbarkeit in der Vielfalt. Bildungsstandards der Kultusministerkonferenz für die Allgemeine Hochschulreife. In J. Kramer, M. Neumann & U. Trautwein (Hrsg.), *Abitur und Matura im Wandel* (S. 29–58). Wiesbaden: Springer.
Statistisches Bundesamt (1967). *Bevölkerung und Kultur. Reihe 10. Bildungswesen. Allgemeinbildende Schulen*. Stuttgart: Kohlhammer.
Steiner, H.-G. (1984). Mathematisch-naturwissenschaftliche Bildung – Kritisch-konstruktive Fragen und Bemerkungen zum Aufruf einiger Fachverbände. In M. Reiß & H.-G. Steiner (Hrsg.), *Mathematikkenntnisse – Leistungsmessung – Studierfähigkeit* (S. 5–58). Köln: Aulis.
Wolff, G. (Hrsg.). (1959). *Die Elemente der Mathematik. Kurzausgabe. Oberstufe. Arithmetik, Algebra, Geometrie, Analysis, Trigonometrie* (5. Aufl.). Paderborn: Schöningh.
Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15–27. https://doi.org/10.1111/j.1745-3992.2010.00190.x.
Wu, M., Tam, H.P., & Jen, T.-H. (2016). *Educational measurement for applied researchers*. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5.
Zieky, M. (1993). DIF statistics in test development. In P.W. Holland & H. Wainer (Hrsg.), *Differential item functioning* (S. 337–347). Hillsdale: Erlbaum.