Rückwirkungen von Large Scale Assessments auf das Bildungssystem

Manfred Prenzel

Der Begriff „Large Scale Assessment“ hat zwar noch nicht in die Alltagssprache Eingang gefunden, wird aber heute in bildungspolitischen und pädagogischen Fachkreisen mit großer Selbstverständlichkeit verwendet. Eigentlich könnte man auch etwas einfacher von „Schulleistungsvergleichen“ sprechen.[1] Diese Bezeichnung kann jedoch Missverständnisse nahelegen, denn oft werden in solchen Studien nicht „Schulleistungen“ in einem engen Sinn erhoben, sondern Facetten von Grundbildung, die in bestimmten Phasen der Bildungsbiographie entwickelt sein sollten. Als Bezugspunkte dienen damit nicht nur nationale Curricula oder Schulfächer (man denke z.B. an Querschnittskompetenzen wie Lesen), sondern theoretisch begründete und international abgestimmte Bildungskonzeptionen. Wenn Kohorten jenseits des schulpflichtigen Alters in solche Untersuchungen einbezogen werden (etwa um den Stand der Bildung über die Lebensspanne zu erfassen), greift die Bezeichnung „Schulleistungsvergleich“ soundso nicht mehr. Die „große Skala“ bringt zum Ausdruck, dass entsprechende modellierte Facetten von Bildung (die Wissen, Fähigkeiten, Fertigkeiten oder Kompetenzen umfassen könnten) an großen, repräsentativen Stichproben erhoben werden, und damit Rückschlüsse auf die Grundgesamtheit erlauben, sei es in nationalen oder internationalen beziehungsweise quer- oder längsschnittlich angelegten Vergleichen.[2] Der Begriff „Assessment“ wiederum verweist auf einen empirischen Erhebungszugang, der auf einer umfassenderen, systematischen Testkonzeption beruht (z.B. mehrere Tests, inzwischen nicht nur Papier- und Bleistift, sondern auch computerbasiert, zusätzliche Erhebungsinstrumente).

Wie im Folgenden ausgeführt wird, haben sich im Verlauf der letzten zwanzig Jahre in Deutschland verschiedene Large Scale Assessments etabliert, die belastbare Befunde und Erkenntnisse über den Bildungsstand bereitstellen. Der erste Abschnitt des Beitrags skizziert zunächst die Ausgangspunkte für diese Bewegung, um dann einen Überblick über die weitere Entwicklung bis heute zu geben. In einem zweiten Schritt werden verschiedene Funktionen solcher Large Scale Assessments betrachtet, die unter dem Aspekt der Rückwirkungen auf das Bildungssystem (bzw. auf Akteure in diesem) interpretiert werden können. Am Ende des Beitrags wird der Versuch unternommen, diese Entwicklungen hinsichtlich ihrer Wirksamkeit einzuschätzen und auf Herausforderungen und Handlungsbedarf hinzuweisen.

1 Rückblick

Die 1995 durchgeführte „Trends in International Mathematics and Science Study“ (TIMSS) war der erste internationale Schulleistungsvergleich, an dem Deutschland nach einer jahrzehntelangen Pause teilgenommen hatte. Die 1997 publizierten Ergebnisse[3] fanden ein beträchtli-

1 *Drechsel/Prenzel/Seidel*, 2014.
2 *Seidel/Prenzel,* 2008.
3 *Baumert et al*., 1997.

RdJB 1/2018, DOI: 10.5771/0034-1312-2018-1-27

ches politisches und öffentliches Interesse. Grund dafür war wohl das Gesamtergebnis, demzufolge die deutschen Schülerinnen und Schüler in den untersuchten Bereichen Mathematik und Naturwissenschaften keineswegs international brillierten, sondern auf dem Durchschnittsniveau der teilnehmenden Staaten platziert waren. Dieser als ernüchternd wahrgenommene Befund galt gleichermaßen für die beiden untersuchten Kohorten auf den Sekundarstufen I und II.

Ein wichtiger Beitrag für die weitere Akzeptanz internationaler Vergleichsstudien kann der Qualität der Testaufgaben und der Testskalierung bei TIMSS zugesprochen werden. Zumindest für die Mathematik und die Naturwissenschaften zeichnete sich ab, dass mit sorgfältig entwickelten Aufgaben nicht nur das Wissen von Fakten und Begriffen zuverlässig abgefragt werden kann, sondern wichtige Aspekte des Verstehens ebenso getestet werden können wie die Fähigkeit, das Wissen in unterschiedlichen Situationen flexibel anwenden zu können. Zusammen mit neuen testtheoretischen Möglichkeiten der Skalierung – auf der Basis von Modellen der Item-Response-Theorie[4]– belegten die Testansätze, dass tatsächlich anspruchsvolle Aspekte einer mathematischen und naturwissenschaftlichen Grundbildung reliabel und valide erhoben werden können.

Während in Deutschland bis zu diesem Zeitpunkt die Vorstellungen von Bildung durch oft abgehobene und abstrakte Charakterisierungen (in der pädagogischen Fachliteratur wie im Feuilleton) in keiner Weise als operationalisierbare, und damit jemals messbare Konstrukte erschienen, waren im internationalen (englischsprachigen) Umfeld seit einigen Jahren pragmatische Konzepte von Grundbildung diskutiert worden, die neue Möglichkeiten der Untersuchung eröffneten. Der (metaphorische genutzte) Begriff der „Literacy“ beschrieb mathematische und naturwissenschaftliche Wissensbestände (deklarativer wie prozeduraler Art), die für die Teilhabe an einer durch Mathematik und Naturwissenschaften geprägten Kultur als notwendig erschienen, die zugleich aber auch als Voraussetzungen für das weiterführende und vertiefende Lernen in diesen Domänen verstanden werden konnten.[5] Für die nachfolgenden Anstrengungen, Bildungsvergleiche mit großen, repräsentativen Stichproben durchzuführen, war dieser Ansatz paradigmatisch: Es ging nun weniger darum, mit diesen Studien Produkte abgeschlossener Bildung zu erfassen, sondern vielmehr entscheidende oder notwendige Voraussetzungen für weiterführende Bildungsprozesse. „Grundbildung“ in Anlehnung an Literacy-Konzepte ließ sich sehr viel besser (und auch für unterschiedliche Phasen schulischer Bildung) theoretisch modellieren und empirisch mit Testverfahren untersuchen.

Die hinter den Erwartungen der Bildungspolitik zurückbleibenden Ergebnisse von TIMSS veranlassten die KMK (1997), sich für eine Teilnahme an dem von der OECD neu konzipierten und langfristig angelegten PISA-Vorhaben zu entscheiden. PISA, das „Programme for International Student Assessment“ startete im Jahr 2000 mit einem (gegenüber TIMSS weiterentwickelten) Testansatz, der neben mathematischer und naturwissenschaftlicher Grundbildung auch Lesekompetenz (im Sinne von Textverstehen) einschloss.[6] Fortan sollten die drei Kompetenzbereiche in einem dreijährigen Turnus regelmäßig bei einer repräsentativen Stichprobe von fünfzehnjährigen Schülerinnen und Schülern untersucht werden, um die internationalen Vergleiche auch über die Zeit (im Sinne eines kontinuierlichen Monitorings) fortführen zu können. Als Vergleichsmaßstab dienen in PISA die in den OECD-Staaten erzielten Ergebnisse.

4 Vgl. *Rost*, 2004.
5 AAAS, 1993, NCTM, 1989.
6 OECD, 1999.

Der erste PISA-Vergleich bekräftigte nicht nur die TIMSS-Befunde, sondern ließ im OECD-Vergleiche weitere zahlreiche und gravierende Probleme Deutschlands im Bildungsbereich erkennen.[7]

Was zeigten die Befunde im Einzelnen? Im Vordergrund stand das Abschneiden im Ranking der OECD. In allen drei untersuchten Bereichen (Lesen, Mathematik, Naturwissenschaften) lag die Leistung der Schülerinnen und Schüler aus Deutschland signifikant unter dem Mittelwert der OECD-Staaten. Für viele erschien dieser Befund als Ausschlag gebend für den so genannten „PISA-Schock". Tatsächlich war es aber ein ganzes Ensemble von Ergebnissen für Deutschland, die den Schock bewirkt haben dürften:

– In allen getesteten Bereichen war das Leistungsniveau niedrig, die Streuung aber (im internationalen Vergleich) sehr hoch;
– Die Leistungen von fast einem Viertel der Jugendlichen lag auf den untersten Kompetenzstufen und weit davon entfernt, den Lehrplananforderungen zu genügen;
– Auf der anderen Seite erreichten relativ wenige Schülerinnen und Schüler Leistungen auf den obersten Kompetenzstufen und wiesen auf eine unzureichende Talentförderung hin;
– In Deutschland war – wie in kaum einem anderen Land – eine sehr starke Kopplung von Herkunft (soziale Lage und Migrationsstatus) und Kompetenz sowie Bildungsbeteiligung festzustellen.

Schockwirkung dürften die Befunde auch deshalb erzielt haben, weil sie in Deutschland weit verbreiteten Überzeugungen widersprachen. So wurde die Annahme widerlegt, dass hohe Leistungen in einem Land notwendig mit einer großen Leistungsstreuung verbunden sein müssten oder Talente und Leistungsspitzen nur in einem gegliederten Schulsystem angemessen gefördert werden könnten. Der internationale Vergleich führte vor, dass die Kopplung von Herkunft und Kompetenz sehr viel niedriger ausgeprägt sein konnte als in Deutschland. Auch wurden durch die Vergleiche bestimmte Mythen tangiert, etwa, dass die Lernleistungen von der Klassengröße abhingen oder verzögerte Schullaufbahnen durch Klassenwiederholungen oder Späteinschulungen ein weltweit selbstverständliches und probates Mittel wären. Eher entstand der Eindruck, dass Deutschland mit der Lebenszeit der Schülerinnen und Schüler großzügig bis fahrlässig umging.

Für weitere Irritationen sorgten dann die Ergebnisse aus systematischen Vergleichen zwischen den Bundesländern, die in Deutschland mit Hilfe von Stichprobenerweiterungen und Zusatzstudien durchgeführt werden konnten.[8] Dabei zeigten sich Leistungsunterschiede zwischen den Bundesländern (in einer Größenordnung bis zu 50 Punkten), die schon fast denen zwischen den gesamten OECD-Staaten nahekamen. Starke Beachtung fanden Befunde über bedeutsame Unterschiede zwischen den Bundesländern in der Kopplung zwischen Herkunft und Kompetenz, die insofern erwartungswidrig waren, als sie oftmals nicht der im jeweiligen Land politisch verkündeten Programmatik entsprachen, besonders für Bildungsgerechtigkeit zu sorgen. Hier zeichnete sich ab, dass Bildungsgerechtigkeit nur sehr bedingt durch (die bisher üblichen) Entscheidungen über die Schulstruktur herbeigeführt werden kann.

PISA, insbesondere auch mit den forschungsorientierten Erweiterungen der Studie in Deutschland, präsentierte eine Fülle von Befunden, die ausgeprägte Problemlagen erkennen ließen und auch tradierte Überzeugungen von Grundprinzipien eines erfolgreichen Bildungs-

7 *Baumert et al*., 2001.
8 *Baumert et al*., 2002.

systems in Frage stellten. Zugleich erschien PISA als leuchtendes Beispiel dafür, dass sich Large Scale Assessments eigneten, um Bilanz zu ziehen, Bildungssysteme zu beobachten, Benchmarks und Anregungen zu ihrer Weiterentwicklung zu erhalten. Eine Rückwirkung von PISA bestand also bereits darin, Large Scale Assessments als unverzichtbares Verfahren für die Qualitätssicherung in Bildungssystemen zu betrachten.

Nachdem PISA speziell den Altersabschnitt gegen Ende der ersten Sekundarstufe abdeckte, lag es nahe, auch am Ende der davorliegenden Bildungsphase Leistungsvergleiche vorzusehen. Für die Primarstufe bot es sich für Deutschland an, wiederum an internationalen Large Scale Assessments teilzunehmen und diese gegebenenfalls zu erweitern (durch Ländervergleiche). So beteiligt sich Deutschland seit 2002 regelmäßig an der „Progress in Reading Literacy Study (PIRLS)“, die in Deutschland unter dem Namen IGLU bekannt ist[9] sowie an der „Trends in Mathematics and Science Study“, einer Nachfolgeuntersuchung zu TIMSS auf der Primarstufe mit dem gleichen Akronym.[10] Large Scale Assessments wurden dann aber auch noch für andere Zielgruppen eingesetzt, etwa seit 2008 bei Lehrkräften für die Mathematik in der „Teacher Education and Development Study in Mathematics /TEDS-M)“.[11] Das seit 2012 verfolgte „Programme for the International Assessment of Adult Competencies” wiederum untersucht seit 2012 in einem 10-Jahres-Takt Kompetenzen in den Bereichen Lesen, Mathematik und technologiebasiertes Problemlösen bei Stichproben im Altersbereich von 16 bis 65 Jahren.[12] In gewisser Weise gibt diese Studie auch Auskunft über die Nachhaltigkeit schulischer Bildungsangebote und macht zugleich auf Qualifizierungs- und Weiterbildungsbedarf aufmerksam. Gerade die letztgenannten Beispiele unterstreichen, dass das Instrument „Large Scale Assessment“ inzwischen im deutschen Bildungssystem als wichtige Informationsquelle genutzt wird und nicht nur zur Qualitätssicherung im Schulbereich dient.

2 Funktionen und Wirkungen von Large Scale Assessment im Bildungssystem

Häufig werden Large Scale Assessments insbesondere zwei Funktionen zugeschrieben: Sie können erstens dem Monitoring dienen und zweitens Benchmarks für die Weiterentwicklung von Bildungssystemen und ihren Einrichtungen bereitstellen.[13]

Während die Funktion des Monitorings mit klaren Anforderungen an die wissenschaftliche Qualität der Datenerhebung und -aufbereitung verbunden ist, bleibt weniger klar, welche Befunde als eine Art von Benchmark verstanden werden können. Insbesondere ist es fraglich, inwieweit bestimmte Benchmarks aufgrund von Befunden in anderen nationalen Bildungssystemen überhaupt sinnvoll als Bezugspunkt oder konkrete Herausforderung in einem anderen nationalen System gewählt werden können. Noch schwieriger wird es, wenn Wege zum Erreichen von Benchmarks in anderen kulturellen Kontexten, Traditionen und unterschiedlich strukturierten und ausgestatteten Bildungssystemen gefunden werden sollen. Die Funktion des Benchmarkings dürfte damit häufig auf einen Anregungscharakter beschränkt sein – und noch keine Maßnahmen begründen können. Allerdings werden Benchmarks aus internationalen oder natio-

9 *Bos et al*., 2003.
10 *Bos et al*., 2008.
11 *Blömeke/Kaiser/Lehmann*, 2010.
12 *Rammstedt et al*., 2013.
13 Vgl. *Seidel/Prenzel*, 2008.

nalen Vergleichen durchaus in der politischen Arena genutzt, um anhand von Beispielen und unter Verweis auf empirische Studien Druck erzeugen zu können.

Betrachtet man die Funktion des Monitorings, dann erscheinen einmalige „Schnappschüsse" der Bildungslandschaft zu einem Zeitpunkt weniger nützlich zu sein als umfassendere Monitoringsysteme, die mit vergleichbaren Indikatoren Beobachtungen über längere Zeiträume ermöglichen und damit eben auch Feedback geben können, nicht nur über Entwicklungen und Trends, sondern vielleicht sogar über die Wirksamkeit von zwischenzeitlich umgesetzten Maßnahmen. Ein elaboriertes Monitoringsystem wird sich nicht nur auf Daten aus einem (z.B. internationalen) Large Scale Assessment stützen, sondern auf mehrere Studien sowie andere Datenquellen aus Bildungsstatistiken nutzen. Es ist durchaus charakteristisch, dass Large Scale Assessment-Programme wie PISA auch im Kontext eines auf Dauer angelegten Berichtssystems stehen, in diesem Fall den jährlichen Berichten der OECD zu „Education at a Glance".[14] In ähnlicher Weise nutzt der Nationale Bildungsbericht in Deutschland[15] soweit möglich und verfügbar Daten aus Large Scale Assessments, um systematisch belastbare Daten zu bestimmten Indikatoren aufzubereiten. Gerade für die Nationale Bildungsberichterstattung in Deutschland gilt, dass diese erst mit Vorliegen von Befunden aus nationalen und internationalen Vergleichsstudien sinnvoll möglich wurde, und damit in gewisser Weise auch als durch Large Scale Assessment angebahnt betrachtet werden kann.

Generell tragen die Anstrengungen, die in Large Scale Assessments vorgenommen werden, um die Qualität von Bildungsergebnissen und -prozessen theoretisch zu modellieren und zu operationalisieren, dazu bei, dass nach normativen Bezugspunkten (und deren Berechtigung) für Vergleiche und Bewertungen von Bildungsprozessen und -ergebnissen gefragt wird. Zweifellos ist es legitim, konkrete Bezugspunkte mit Verweis auf eine curriculare Verankerung angestrebter Bildungsergebnisse zu wählen. Aber auch theoretisch begründete Kompetenzmodelle (eventuell verbunden mit Evidenz über eine mehr oder weniger vorbereitete Anschlussfähigkeit) könnten aufgegriffen werden, und wiederum Diskussionen über die Qualität und Aktualität von Curricula stimulieren wie auch über mehr oder weniger verengte (z.B. eindimensionale) Vorstellungen von Bildung. Interessanter sind vielleicht noch Rückwirkungen auf eine präzisere Bestimmung von übergeordneten (zum Teil verfassungsmäßig begründeten) Bezugspunkten wie zum Beispiel Ansprüchen an „Bildungsgerechtigkeit" (z.B. gleiche Chance unabhängig von Geschlecht, sozialer, ethnischer oder regionaler Herkunft), vor dem Hintergrund von Befunden über Disparitäten, auch bei Kontrolle von bestimmten Merkmalen wie getestete Leistungsfähigkeit und kognitive Grundfähigkeiten.

Die für Deutschland drastischen Befunde aus den ersten internationalen und nationalen Vergleichen im Rahmen von PISA haben die Bildungspolitik erheblich unter Druck gesetzt, Maßnahmen zu bedenken und zu ergreifen, die möglichst bald zu einer Verbesserung der Kompetenzen und zur Lösung von gravierenden Problemen (ungleiche Bildungschancen zum Beispiel) beitragen können sollten. Allerdings wiesen viele Befunde aus diesen Studien darauf hin, dass Unterschiede in den Kompetenzen oder Disparitäten nicht einfach (bzw. kausal) durch strukturelle Merkmale (etwa Grundstruktur des Schulsystems) erklärt werden können. Damit wurde die politische Suche (im üblichen Instrumentenkasten) nach Interventionen auf struktureller Ebene erheblich beeinträchtigt, denn für viele Optionen lag aus den aktuellen Studien keine ausreichende Evidenz vor. Vielmehr deuteten zahlreiche Befunde (in Übereinstimmung

14 Z.B. OECD, 2017.
15 Z.B. Autorengruppe Bildungsberichterstattung, 2016.

mit anderen Studien) darauf hin, dass lernprozessferne Merkmale des Schulsystems weniger Effekt haben als lernprozessnahe Bedingungen (wie zum Beispiel die Qualität des Unterrichts). Bereits die ersten Large Scale Assessments machten darauf aufmerksam, dass das Lehren und Lernen in Deutschland (zu) wenig kumulativ angelegt ist, Schwächen (zum Beispiel in der Lesekompetenz) von Lehrkräften erst zu spät oder nicht adäquat diagnostiziert werden, und insgesamt (gerade auch im internationalen Vergleich) wenig Möglichkeiten zur Qualitätssicherung genutzt werden.

Betrachtet man die von KMK (2002) beschriebenen zentralen Handlungsfelder nach PISA, dann ist zu spüren, dass vielmehr Ansatzpunkte gesucht und gefunden wurden, die Aussichten auf beträchtliche Lernprozesswirkungen boten und zugleich mit Vorstellungen von umsetzbaren Maßnahmenpaketen jenseits von simplen Struktureingriffen verbunden waren. Folgende Felder hatte die KMK (2002) strategisch herausgehoben:

(a) Verbesserung der Sprachkompetenz,
(b) Bessere Verzahnung von Vor- / Grundschule,
(c) Kompetenz Lesen, mathematisches und naturwissenschaftliches Verständnis verbessern,
(d) Förderung bildungsbenachteiligter Kinder,
(e) Qualitätssicherung mit Bildungsstandards,
(f) Professionelle Qualität der Lehrertätigkeit,
(g) Ausbau Ganztagsangebote.

Einige dieser Handlungsfelder (z.B. c und f) verweisen auf Erkenntnisse, die bereits im Anschluss an TIMSS und damit verbundene Videostudien des Mathematikunterrichts sowie aufgrund von Evidenz aus der domänenspezifischen Lehr-Lern- sowie Professionalisierungsforschung in Gutachten[16] zusammengefasst worden waren. Aufgrund der bis dahin erfolgten Umsetzung in Modellprogramm zur Qualitätsentwicklung erschienen solche Ansätze als erfolgversprechend.[17] An dieser Stelle kann ergänzt werden, dass Studien wie TIMSS und PISA in Deutschland auch den Boden vorbereitet haben für großflächige Bemühungen, den Mathematik- und Naturwissenschaftsunterricht mit Hilfe von Qualitätsentwicklungs- und Professionalisierungsprogrammen[18] zu verbessern. Um Übrigen wurden diese Programme wiederum unter Verwendung von Large Scale Assessment-Samples (als eine Art nationaler Kontrollgruppe) einer Evaluation unterzogen wurden, die durchaus positive Effekte auf Unterrichtsqualität und Lernerfolg bestätigte.[19]

Sehr starke Rückwirkungen von Large Scale Assessments auf das Bildungssystem lassen sich exemplarisch (und besonders deutlich) bezogen auf Überlegungen zeigen, die auf die Einführung von Bildungsstandards zielten (Handlungsfeld (e) der KMK). Befunde der ersten PISA-Runde (die durch andere Studien bekräftigt wurden) über die Verteilungen von Schülerinnen auf Kompetenzstufen dürften dafür wohl der entscheidende Anlass gewesen sein: Fast ein Viertel der bei PISA getesteten Fünfzehnjährigen erreichte ja in Lesen, Mathematik und Naturwissenschaften nur ein Kompetenzniveau, das mehr oder weniger den Anforderungen entsprach, die bereits am Ende der Grundschulzeit gemeistert werden sollten. Die in Deutschland üblichen gewichtigen und detaillierten Lehrpläne schienen nicht vor Misserfolgen bei der Ziel-

16 Z.B. BLK, 1997.
17 *Prenzel*, 2000.
18 Z.B. SINUS, vgl. *Prenzel*; *Friedrich/Stadler*, 2009.
19 *Dalehefte et al*., 2014, 2015; *Prenzel et al*., 2005, *Rieck et al*., 2015.

erreichung zu schützen. Es war auch nicht auszuschließen, dass sie in ihrer Detailliertheit vielleicht sogar hinderlich waren, den tatsächlichen Stand des Könnens von Schülerinnen und Schüler festzustellen oder zu diagnostizieren. Zumindest ließ der internationale Vergleich durchaus erfolgreiche Staaten erkennen, die mit schlanken (und anders, nämlich kompetenz- oder outcome-orientiert angelegten) Curricula arbeiteten. Weitere empirische Beobachtungen – wie etwa der beträchtlichen Unterschiede zwischen den Bundesländern in den Kompetenzverteilungen oder in den Zusammenhängen mit sozialer Herkunft – regten dazu an, über möglichst von allen Schülerinnen und Schülern zu erreichende Kompetenzniveaus und deren Verankerung als nationaler Referenz nachzudenken. Eine gründliche Auseinandersetzung mit verschiedenen Möglichkeiten, durch Bildungsstandards Probleme im deutschen Schulsystem beheben zu können, erfolgte in einem umfassenden Gutachten.[20] In dieser Expertise wurde ein Konzept von Bildungsstandards vorgeschlagen, das sich durch eine hohe Fokussierung auf wesentliche Kompetenzen im Sinne von domänenspezifischer Grundbildung auszeichnete. Diese Kompetenzen sollten wiederum anhand von Beispielaufgaben veranschaulicht werden. Wesentlich war die Forderung nach theoretisch begründeten Kompetenzmodellen zur Grundlegung von Standards. Auf diese Weise wurde eine wichtige Voraussetzung erfüllt, um das Erreichen von Standards mit Hilfe von Testverfahren valide erfassen zu können.

Auf der Grundlage dieser Expertise boten sich mehrere Perspektiven für eine verbesserte Qualitätssicherung an: Durch Aufgaben konkretisierte (und theoretisch strukturierte) Bildungsstandards konnten die Möglichkeiten von Lehrkräften erweitern, den Stand des Wissens und Könnens ihrer Schülerinnen und Schüler besser zu erkennen, auch um frühzeitig mit Fördermaßnahme anzusetzen. Im Rahmen von landes- oder bundesweiten Vergleichsarbeiten, für die standardbezogene Testaufgaben entwickelt werden sollten, konnten den Lehrkräfte eine zusätzliche Möglichkeit verschafft werden, den Kompetenzstand ihrer Klasse einzuschätzen und darauf pädagogisch zu reagieren, gegebenenfalls (oder möglichst) auch im Rahmen gemeinsamer Qualitätsentwicklungsmaßnahmen auf der Ebene der jeweiligen Schule. Schließlich bestand noch die weiterreichende Möglichkeit, mit standardbezogenen Tests deutschlandweit zu bestimmten Abschnitten der Schullaufbahn der Frage nachzugehen, inwieweit die Schülerinnen und Schüler auf die Anforderungen (etwa des mittleren Abschlusses) vorbereitet sind. Entsprechende Untersuchungen können die Vergleiche zwischen Bundesländer durch Vergleiche über die Zeit bereichern und wiederum wichtige Erkenntnisse (oder Feedback) über den aktuellen Bildungsstand und Trends liefern. Viele der Vorstellungen zu Bildungsstandards, die in dem Gutachten und in voraus- und nachlaufenden Gesprächen erörtert worden waren, wurden von der Bildungspolitik zügig umgesetzt: Mit der Einrichtung des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) wurde 2004 die institutionelle Grundlage für die Qualitätsentwicklung mit Hilfe von nationalen Bildungsstandards geschaffen. Auf dieser Basis konnten für wichtige Bereiche und Schulstufen Bildungsstandards ausgearbeitet und verbindlich verabschiedet werden und es konnten Aufgabenpools entwickelt, Vergleichsarbeiten eingeführt und nationale Vergleichstests durchgeführt werden. Die nationalen Vergleichsstudien, die mit repräsentativen Stichproben und vergleichbaren Aufgabensamples das Erreichen der Bildungsstandards regelmäßig untersuchen[21], sind als typische Large Scale Assessments einzuordnen, die mit einem etwas anderen Design und Bezugssystem sowie mit anders akzentuierten Fragestellungen die internationalen Vergleichsstudien sinnvoll ergänzen.

20 *Klieme et al*., 2003.
21 Vgl. zuletzt *Stanat et al*., 2017.

Betrachtet man die relativ junge Geschichte von Large Scale Assessments in Deutschland, dann scheint der Rückwirkungseffekt dieser Studien auch davon abzuhängen, inwieweit die (vor allem international) auf Kompetenzmessung fokussierten Studien in ihrem Design erweitert wurden, um weiterführende Forschungsfragen zu beantworten. So war es für die Akzeptanz der Befunde aus der ersten PISA-Erhebung durchaus hilfreich, dass die Testaufgaben hinsichtlich ihrer Übereinstimmung mit Lehrplananforderungen in Deutschland abgeglichen wurden. So zeigten Zusatzstudien (an einem zweiten Testtag) mit Mathematikaufgaben, die eng auf die Lehrpläne in Deutschland und hiesige Traditionen der Aufgabenstellung bezogen waren, dass die Ergebnisse auch unter Verwendung „typisch deutscher" Testaufgaben nicht besser ausgefallen wären. Weiterführende Erkenntnisse über die Nachhaltigkeit des Lehrens und Lernens in Deutschland brachten Follow-up-Studien, die ein Jahr nach der PISA-Erhebung die weitere Kompetenzentwicklung mit Testbatterien erfassten.[22] Beide Studien belegten, dass bei einem erheblichen Anteil der Schülerinnen und Schülern von der 9. zur 10. Klassenstufe keine nennenswerten Kompetenzzuwächse zu verzeichnen sind – wiederum ein Befund, der auf unzureichend kumulatives Lernen hinweist. Eine bemerkenswerte Rückwirkung ging schließlich von einer ebenfalls an PISA angekoppelten Studie aus. Das an PISA 2003 angekoppelte Follow-up-Design wurde in der sogenannten COACTIV-Studie[23] genutzt, um die Mathematiklehrkräfte, die in den PISA-Klassen unterrichteten, mit einem umfassenden Assessment zu untersuchen. Dies war der erste Versuch in Deutschland, auch Komponenten des professionellen Wissens von Lehrkräften (einschließlich Fachwissen) mit Hilfe von Testverfahren im Rahmen eines Large Scale Assessments zu erheben und auf die Lernfortschritte der von ihnen unterrichteten Schülerinnen und Schüler zu beziehen. Zusammen mit großangelegten Videountersuchungen von Unterrichtsstunden[24] trugen COACTIV und nachfolgende Studien dazu bei, Lehrer- und Unterrichtsmerkmale zu identifizieren, die für die Kompetenz- und auch Motivationsentwicklung der Schülerinnen und Schüler prädiktiv waren.

3 Bilanz und Ausblick

Man kann sich fragen, wie sich das Bildungssystem in Deutschland ohne Teilnahme an TIMSS, PISA und anderen Large Scale Assessments weiterentwickelt hätte. Wie lange hätte sich die Vorstellung halten können, dass das System, so wie es ist, gut und kaum zu verbessern ist? Wäre es denkbar gewesen, längerfristig internationale Vergleiche und dort identifizierte starke Bildungssysteme zu ignorieren oder abzutun? Und könnte man sich heute aus solchen Vergleichen einfach ausklinken?

So lange die theoretische und methodische Qualität von Large Scale Assessments dem aktuellen Stand der Forschung gerecht wird, dürfte es kaum mehr einem Land möglich sein, sich solcher Vergleichsstudien zu entziehen. Dabei darf man unterstellen, dass dieser Qualitätsanspruch auch dazu führt, die Testkonzeptionen und Methoden ständig weiterzuentwickeln – immer unter der Zielsetzung, relevante Bildungsergebnisse und ihre Bedingungen valide und reliabel zu erfassen.

22 Z.B. *Prenzel et al*., 2006; *Reiss et al*., 2017.
23 Z.B. *Kunter et al*., 2011.
24 Z.B. *Seidel et al.,* 2006.

Gewicht erhalten die Ergebnisse solcher Studien, wenn sie auf wesentliche Anforderungen in Bildungssystemen bezogen sind, also etwa auf Zielsetzungen in Lehrplänen, auf Bildungsstandards und auf übergeordnete Ansprüche gleicher bzw. gerechter Bildungschancen. Hier besteht eine wichtige Funktion von Large Scale Assessments darin, Stärken, Schwächen und mögliche Probleme zu identifizieren. Mit Hilfe von Large Scale Assessments können auch unrealistische oder überholte Anforderungen, z.B. in Lehrplänen ausgemacht werden. Large Scale Assessments können Eindrücke – etwa von einer unzureichenden Ausbildungsfähigkeit von Jugendlichen oder von Niveauverlusten aufgrund der Bildungsexpansion – bekräftigen oder relativieren und widerlegen.

Denkt man an die Jahre zurück, in denen politisch über eine Teilnahme an Large Scale Assessments diskutiert und entschieden wurde, dann waren damals von verschiedenen Seiten sehr wohl bereits durchaus kritische Einschätzungen der Qualität der Schulen und Bildungsergebnisse in Deutschland zu vernehmen. Auch aus diesen Gründen lag es nahe, sich einmal internationalen Vergleichen zu stellen. Mit Large Scale Assessments wurde zudem deutlich, dass über Befragungen, Tests und systematisch erhobene Daten viele interessante und steuerungsrelevante Informationen in Erfahrung gebracht werden können, die bisher nicht im Blick waren, etwa über die (zwischen Ländern stark variierenden) Anteile von verzögerten Schullaufbahnen oder über die Bildungschancen von Jugendlichen mit Zuwanderungshintergrund, und zwar nicht einfach bestimmt durch die Staatsbürgerschaft, sondern durch das Geburtsland der Eltern und das Geburtsland des Kindes. Die Differenzierung von erster und zweiter Generation der Zuwanderung wurde erst über PISA bekannt und nachfolgend ein wichtiger Aspekt des weiteren Monitorings.

So kann man für die junge Geschichte der Large Scale Assessments im Bildungsbereich feststellen, dass diese wesentlich zu einer „empirische Wende“[25] in der Bildungsadministration und -politik, aber auch in der Öffentlichkeit beigetragen haben. Als Beleg dafür mag die Gesamtstrategie der KMK (2015) dienen, die eine Teilnahme an internationalen Vergleichsstudien, die Qualitätsentwicklung mit Hilfe von Bildungsstandards und die Bildungsberichterstattung als wesentliche Säulen enthält und institutionell verankert hat. Large Scale Assessments haben sich offensichtlich aus bildungspolitischer Sicht als so nützlich und relevant erwiesen, dass auch in Zukunft wesentliche Beiträge von ihnen erwartet werden.

Für eine Bildungspolitik, die Evidenz (im Sinne des besten verfügbaren Wissens) als Referenz für Entscheidungsfindungen wünscht, sind Large Scale Assessments ein wichtiges Instrument geworden. Das heißt allerdings nicht, dass die Bildungspolitik die Erkenntnisse, die von Large Scale Assessments bereit gestellt werden, immer schon für ihre Zwecke und Steuerungsanliegen als ausreichend betrachtet. Gerade wenn Large Scale Assessments Probleme identifizieren, liegt es nahe, dass politische Akteure sich zugleich oder möglichst schnell Evidenz wünschen, die Erfolgsaussichten von Maßnahmen zur Problemlösung einschätzen lässt. Diese Erwartung beruht vermutlich auf einem Missverständnis über die Art von Wissen, die durch Large Scale Assessments bereitgestellt wird.[26] Diese Ansätze gelangen selbstverständlich an Grenzen, wenn (kausale) Erklärungen für bestimmte Zustände oder Problemlagen erwartet werden. Und die Hoffnung auf empirisch gesichertes Veränderungswissen (bezogen auf in der Situation anwendbare zielbezogene Maßnahmen) muss gänzlich enttäuscht werden, denn auch dazu würden andere Forschungsdesigns (und theoretische Grundlagen) benötigt. Diese Miss-

25 *Lange*, 2008.
26 *Bromme/Prenzel/Jäger*, 2014.

verständnisse weisen darauf hin, dass es bis jetzt wohl noch nicht ausreichend gelungen ist, die Möglichkeiten und Grenzen von Large Scale Assessments zu kommunizieren. Vielleicht liegt auch dies an der jungen Geschichte dieser Zugänge im Bildungskontext, denn im Gesundheitsbereich dürfte es heutzutage kaum mehr vorkommen, dass Politik und Öffentlichkeit zum Beispiel von epidemiologischen Studien sich Antworten darauf erhoffen, wie etwa Krebs- oder Demenzerkrankungen kuriert werden können.

Large Scale Assessments haben in Deutschland nicht nur Problemlagen sichtbar gemacht, sondern jeweils auch Forschungsbedarfe festgestellt. Seither wurden und werden in einer ganzen Reihe von Programmen Forschungsfragen bearbeitet, die zum Teil sehr eng mit Large Scale Assessment zu tun haben, etwa in den DFG-Schwerpunktprogrammen zu Bildungsqualität von Schule[27] und zur Kompetenzmodellierung[28] oder im Rahmenprogramm des BMBF zur Empirischen Bildungsforschung.[29] Gerade solche Programme können geeignet sein, ein breites Spektrum von empirischen Studien zu fördern, die vertiefende Erkenntnisse zu den Befunden aus Large Scale Assessments beitragen, Erklärungen beisteuern oder Maßnahmen erproben. Bestimmte Studien können aber auch helfen, die Erhebungs- und Auswertungsverfahren von Large Scale Assessments weiterzuentwickeln. Besonders hervorzuheben ist auch die Einrichtung einer National Educational Panel Study (NEPS), die letztlich Large Scale Assessments in einem longitudinalen und Panel-Design über die gesamte Lebensspanne systematisch integriert[30] – eine Studie, die nicht nur zum besseren Verständnis von Bedingungen der Kompetenzentwicklung beiträgt, sondern auch die Qualität von Vorhersagen aus Kompetenztests für die weitere Biographie besser abschätzen lässt.

Herausforderungen für künftige Large Scale Assessments gibt es zahlreiche. Kleinerer Art sind vermutlich jene, die sich auf Aspekte der Durchführung (etwa als computerbasierte Assessments) richten. Größere Herausforderungen dürften darin bestehen, das Spektrum von Bildungsergebnissen breiter auszuleuchten, um die inhaltliche Breite von Kompetenzen besser abzudecken, vor allem aber, um der Mehrdimensionalität von Bildungszielen gerecht zu werden. Vor allem wären hier Versuche zu unternehmen, Merkmale – angefangen von Motivation und Interesse bis zu Wertorientierungen und sozialen Kompetenzen – nicht nur mit Selbsteinschätzungsverfahren zu erfassen, sondern mit aussagekräftigen Tests. Auch Merkmale von Schulen werden bislang mit Fragebögen (aus der Perspektive der Schulleitung) erfasst; hier könnten Assessments zu stärker belastbaren Befunden führen. Schließlich könnten Merkmale, die für die Professionalität von Lehrkräften relevant sind, im größeren Umfang für Large Scale Assessments erschlossen werden, und nicht zuletzt der Unterricht selbst.

Am Ende kann die Frage gestellt werden, ob die bisherigen Large Scale Assessments in Deutschland selbst zu der Verbesserung der Bildungsergebnisse beigetragen haben, die zwischen 2000 und 2015 zum Beispiel bei PISA beobachtet wurden.[31] Nachdem die Problemlagen wie die Fortschritte mit Large Scale Assessments erfasst wurden, könnte vermutet werden, dass sich Unterricht und Schulen in Deutschland stärker auf Anforderungen dieser Art von Tests hin ausgerichtet haben. Dagegen spricht die Tatsache, dass PISA nach wie vor ein Test ohne Konsequenzen für die Schülerinnen und Schüler wie auch Lehrkräfte und Schulen ist („low stake

27 *Prenzel/Allolio-Näcke*, 2006.
28 *Leutner et al.*, 2017.
29 *Buchhaas-Birkholz*, 2009; BMBF, 2017.
30 *Blossfeld/Roßbach/von Maurice*, 2011.
31 *Reiss et al*., 2016.

assessment“, an einer Stichprobe) – ein Alignment in Richtung Assessment dürfte eher für „high stake“-Tests erwartet werden. Dagegen sprechen ebenso Befunde aus Versuchen, die zeigen, dass das Lösen von PISA-Aufgaben nicht einfach durch ein Testtraining oder Coaching verbessert werden kann.[32] Gegen einen simplen Testeffekt spricht auch die Variabilität – man könnte auch sagen: die Fragilität – der beobachteten Leistungszuwächse. Sowohl PISA 2015[33] als auch der jüngste Bildungstrend[34] liefern einige Hinweise dafür, dass die positive Entwicklung zum Teil ins Stocken geraten sein könnte.

Insgesamt ist es äußerst schwierig, die verschiedenen Einflussfaktoren zu bestimmen, die seit PISA 2000 dazu beigetragen haben können, dass sich das Leistungsniveau verbessert hat und Disparitäten geringer wurden. Entsprechende Betrachtungen finden sich in den Schlusskapiteln aller deutschen PISA-Berichte. Jedoch dürften sich bestimmte Vermutungen nicht von der Hand weisen lassen, etwa die, dass durch Large Scale Assessments Schule, Bildung und die Frage nach der Leistungsfähigkeit in Deutschland mehr Aufmerksamkeit erfahren haben. Ebenso spricht viel dafür, dass in den oben genannten strategischen Handlungsfeldern der KMK (2002) sehr viel unternommen wurde, insbesondere in den Bereichen Qualitätssicherung, Qualitätsentwicklung und Professionalisierung. Dass Large Scale Assessments hier sehr viele indirekte, aber nennenswerte Effekte hatten, wurde an anderen Stellen[35] und in diesem Beitrag anhand von Beispielen erläutert. Man darf gespannt sein, welche Ergebnisse die nächsten Large Scale Assessments liefern werden.

Literatur

AAAS (American Association for the Advancement of Science) (1993). *Benchmarks for science literacy. Project 2061.* New York: Oxford University Press.

Autorengruppe Bildungsberichterstattung (2016). *Bildung in Deutschland 2016.* Bielefeld: W. Bertelsmann Verlag.

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

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Tillmann, J. & Weiß, M. (Hrsg.) (2002). *PISA 2000. Die Länder der Bundesrepublik Deutschland im Vergleich.* Opladen: Leske & Budrich.

Baumert, J., Lehmann, R. et al. (1997). *TIMSS — Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde.* Opladen: Leske & Budrich.

BLK. Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung (Hrsg.) (1997). *Gutachten zur Vorbereitung des Programms „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts”.* Materialien zur Bildungsplanung und zur Forschungsförderung, Heft 60. Bonn: BLK.

Blömeke, S., Kaiser, G., & Lehmann, R. (Hrsg.) (2010). *TEDS-M 2008 - Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich.* Münster: Waxmann.

Blossfeld, H.-P., Roßbach, H.-G. & von Maurice, J. (Eds.) (2011). Education as a lifelong process. The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft, Sonderheft 14*.

BMBF (2017). *Rahmenprogram Empirische Bildungsforschung des Bundesministeriums für Bildung und Forschung.* Berlin: BMBF.

http://www.empirische-bildungsforschung-bmbf.de/media/content/Rahmenprogramm%20empirische%20Bildungsforschung%20(BITV).pdf (zuletzt abgerufen am 19.4.2018).

32 *Brunner et al*., 2007.
33 *Reiss et al*., 2016.
34 *Stanat et al*., 2017.
35 *Prenzel/Blum/Klieme,* 2015, *Sälzer/Prenzel*, 2017.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Walther, G. & Valtin, R. (Hrsg.) (2003). *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich.* Münster: Waxmann.

Bos, W., Bonsen, M., Baumert, J., Prenzel, M., Selter, C. & Walther, G. (Hrsg.) (2008). *TIMSS 2007. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich.* Münster: Waxmann.

Bromme, R. Prenzel, M. & Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. Eine Analyse von Anforderungen an die Darstellung, Interpretation und Rezeption empirischer Befunde. *Zeitschrift für Erziehungswissenschaft, Sonderheft 27*, 3-54.

Brunner, M., Artelt, C., Krauss, S. & Baumert, J. (2007). Coaching for the PISA test. *Learning and Instruction 17* (2), 111-122.

Buchhaas-Birkholz, D. (2009). Die „empirische Wende" in der Bildungspolitik und in der Bildungsforschung. Zum Paradigmenwechsel des BMBF im Bereich der Forschungsförderung. *Erziehungswissenschaft 20 (Heft 39)*, 27-33.

Dalehefte, I.-M., Wendt, H., Köller, O., Wagner, H., Pietsch, M., Fischer, C., & Bos, W. (2014). Bilanz von neun Jahren SINUS an deutschen Grundschulen: Evaluation im Rahmen der TIMSS 2011 - Erhebung. *Zeitschrift für Pädagogik 60* (2), 245–263.

Dalehefte, I. M., Rieck, K., Wendt, H., Kasper, D., Köller, O. & Bos, W. (2015). Mathematische Kompetenzen von Lernenden aus SINUS-Grundschulen im Vergleich zu TIMSS 2011. In H. Wendt, T. C. Stubbe, K. Schwippert & W. Bos (Hrsg.), *10 Jahre international vergleichende Schulleistungsforschung in der Grundschule. Vertiefende Analysen zu IGLU und TIMSS 2001 bis 2011* (S. 185–200). Münster: Waxmann.

Drechsel, B., Prenzel, M. & Seidel, T. (2014). Nationale und internationale Schulleistungsstudien. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie. 2. Auflage* (S. 343- 368). Heidelberg: Springer.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise.* Bonn: BMBF.

KMK (1997). Grundsätzliche Überlegungen zu Leistungsvergleichen innerhalb der Bundesrepublik Deutschland – Konstanzer Beschluss**.** *Beschluss der Kultusministerkonferenz vom 24.10.1997.*

KMK (2002). *PISA 2000 – Zentrale Handlungsfelder. Zusammenfassende Darstellung der laufenden und geplanten Maßnahmen in den Ländern*. Bonn: KMK. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2002/2002_10_07-Pisa-2000-Zentrale-Handlungsfelder.pdf (zuletzt abgerufen am 19.4.2018).

KMK (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Bonn: KMK. http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf (zuletzt abgerufen am 19.4.2018).

Kunter, M., Baumert, J., Blum, W., Klusmann, U. Krauss, S. Neubrand, M. (Hrsg). (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Lange, H. (2008). Vom Messen zum Handeln: "Empirische Wende" der Bildungspolitik? *Recht der Jugend und des Bildungswesens, 56*(1), 7-14.

Leutner, D., Fleischer, J., Grünkorn, J. & Klieme, E. (Eds.) (2017). *Competence assessment in education. Research, models and instruments*. New York: Springer.

NCTM (National Council of Teachers of Mathematics) (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: NCTM.

OECD (1999). *Measuring student knowledge and skills. A new framework for assessment.* Paris: Organisation for Economic Co-Operation and Development.

OECD (2017). *Bildung auf einen Blick 2017. OECD Indikatoren.* Bielefeld: W. Bertelsmann Verlag.

Prenzel, M. (2000). Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts: Ein Modellversuchsprogramm von Bund und Ländern. *Unterrichtswissenschaft, 28,* 103-126.

Prenzel, M. & Allolio-Näcke, L. (Hrsg.) (2006). *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms.* Münster: Waxmann.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rost, J. & Schiefele, U. (Hrsg.) (2006). *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres.* Münster: Waxmann.

Prenzel, M. Blum, W. & Klieme, E. (2015). The impact of PISA on mathematics teaching and learning in Germany. In K. Stacey & R. Turner (Eds.), *Assessing mathematical literacy. The PISA experience* (pp. 239-248). Heidelberg/NewYork: Springer.

Prenzel, M., Carstensen, C. H., Senkbeil, M., Ostermeier, C. & Seidel, T. (2005). Wie schneiden SINUS-Schulen bei PISA ab? Ergebnisse der Evaluation eines Modellversuchsprogramms. *Zeitschrift für Erziehungswissenschaft,* 8 (4), 487-501.

Prenzel, M., Friedrich, A. & Stadler, M. (Hrsg.) (2009). *Von SINUS lernen – Wie Unterrichtsentwicklung gelingt.* Seelze-Velber: Klett/Kallmeyer.

Rammstedt, B., Ackermann, D., Helmschrott, S., Klaukien, A., Maehler, D.B., Martin, S., Massing, N. & Zabal, A. (2013). *Grundlegende Kompetenzen Erwachsener im internationalen Vergleich. Ergebnisse von PIAAC 2012.* Waxmann, Münster.

Reiss, K., Klieme, E., Köller, O. & Stanat, P. (Hrsg.) (2017). PISA plus 2012 - 2013. *Zeitschrift für Erziehungswissenschaft*, Sonderheft 33.

Reiss, K., Sälzer, Ch., Schiepe-Tiska, A., Klieme, E. & Köller, O. (Hrsg.) (2016). *PISA 2015. Eine Studie zwischen Kontinuität und Innovation.* Münster/New York: Waxmann.

Rieck, K., Dalehefte, I.M., Wendt, H. & Kasper, D. (2015). Wie schneidet das Unterrichtsentwicklungsprogramm SINUS an Grundschulen im Vergleich zu TIMSS 2011 ab? Evaluation der Naturwissenschaftsbezogenen Daten. *Zeitschrift für Grundschulforschung,* 8 (1), 39-52.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion.* Bern: Huber.

Sälzer, Ch. & Prenzel, M. (2017). Policy implications of PISA in Germany: The case of teacher education. In L. Volante (Ed.). *The PISA effect on global educational governance* (pp. 109 – 125). Oxford: Routledge.

Seidel, T., Prenzel, M., Rimmele, R., Schwindt, K., Kobarg, M., Herweg, C. & Dalehefte, I. M. (2006). Unterrichtsmuster und ihre Wirkungen. Eine Videostudie im Physikunterricht. In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms.* (S. 99-123). Münster: Waxmann.

Seidel, T. & Prenzel, M. (2008). Large scale assessment. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts. State of the art and future prospects* (p. 279-304). Göttingen: Hogrefe & Huber.

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S., & Haag, N. (Hrsg.) (2017). *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich*. Münster: Waxmann.

Verf.: Prof. Dr. Manfred Prenzel, Leiter des Zentrums für LehrerInnenbildung, Universität Wien, Porzellangasse 4, A-1090 Wien, E-Mail: manfred.prenzel@univie.ac.at