1. Empirische Forschungen und Leistungserhebungen im Mathematikunterricht[1]

Hans-Dieter Sill

1.1. Vorbemerkungen

Wir verwenden den Begriff **Leistungserhebung** als Oberbegriff für sämtliche Verfahren zur Ermittlung von Schülerleistungen im Unterricht. Das Spektrum der Leistungserhebungen reicht von der Kurzkontrolle eines einzelnen Lehrers über das Erfassen der Leistungen aller Schüler eines Landes bis zu internationalen Vergleichsstudien. Als Leistung eines Schülers verstehen wir nach Heller; Hany (2001, S. 88) ein in einer Anforderungssituation erzeugtes nachweisbares Tätigkeitsprodukt eines Schülers, für dessen Bewertung es Gütemaßstäbe gibt.

Für bestimmte Typen von Leistungserhebungen sind spezielle Bezeichnungen üblich, wie Klassenarbeiten, Parallelarbeiten, informelle und standardisierte Schulleistungstests u. a. Der Terminus Vergleichsarbeiten wird unterschiedlich verwendet. Helmke und Hosenfeld (2003) verstehen darunter schriftliche Arbeiten, die in einer größeren Anzahl von Schulen auf der Basis einer vorgegebenen Aufgabenstichprobe eingesetzt werden mit dem Ziel, die Leistungen der Schüler an einer klassen- und schulübergreifenden sozialen und/oder kriterialen Bezugsnorm zu messen. Wir sind der Meinung, dass es für die Kommunikation auf nationaler Ebene besser wäre, auf solche Wortschöpfungen wie Lernstandserhebungen, Orientierungsarbeiten, Diagnosearbeiten oder Kompetenztests zu verzichten und nur von Leistungserhebungen zu sprechen, deren spezielle Eigenschaften durch zusätzliche wörtliche Beschreibungen angegeben werden. Außer den weitgehend einheitlich verwendeten Bezeichnungen wie Kurzarbeit, Klausur, Abschlussprüfung und psychologischer Test werden spezielle Bezeichnungen von uns nur in Verbindung mit konkreten Projekten benutzt.

Leistungserhebungen gehören zur Arbeit eines jeden Lehrers, sie können aber auch ein Instrument oder sogar Ziel einer wissenschaftlichen Untersuchung sein. In diesem Buch geht es vor allem um Leistungserhebungen im Rahmen wissenschaftlicher Untersuchungen und um zentrale Leistungserhebungen von Schulbehörden. Ein Vergleich ihrer Funktionen, Ziele und Merkmale mit den üblichen Leistungserhebungen in der Schule kann aber zum tieferen Verständnis ihres Anliegens und Nutzens beitragen, wie im Kapitel 2 nachgewiesen werden soll. …

2. Merkmale und Funktionen von Leistungserhebungen

Hans-Dieter Sill

2.1. Merkmale von Leistungserhebungen

Bevor ein System von Funktionen von Leistungserhebungen vorgestellt und diskutiert wird, sollen Aspekte und Merkmale von Leistungserhebungen zusammengestellt und erläutert werden. Die Merkmale werden im folgenden Text **fett** hervorgehoben.

Zunächst wollen wir jedoch einige generelle Bemerkungen zu unseren Auffassungen über die Entstehung und Erfassung statistischer Daten bei Leistungserhebungen machen.

Im Zusammenhang mit einer Leistungserhebung müssen zwei Prozesse unterschieden werden,

---

[1] Auszug aus Sill, H.-D.; Sikora, Ch. (2007): Leistungserhebungen im Mathematikunterricht: Theoretische und empirische Studien. Hildesheim, Franzbecker, S. 10, 48-81

- der Prozess der Entwicklung der Personenmerkmale, die mit der Leistungserhebungen erfasst werden sollen, sowie
- der Prozess der Erfassung der Daten, d. h. die Leistungserhebungen selbst.

Wir fassen diese Prozesse als zufälliger Vorgänge auf, die unter bestimmten Bedingungen (Einflussfaktoren) ablaufen und mehrere mögliche Ergebnisse besitzen, die mit einer bestimmten Wahrscheinlichkeit eintreten können. Die Wahrscheinlichkeit der Ergebnisse hängt von den Bedingungen der Vorgänge ab. Die Bedingungen sind in der Regel wieder Ergebnisse anderer zufälliger Vorgänge.

Soll etwa mit einem Test das Wissen und Können eines Schülers im Lösen von linearen Gleichungen erfasst werden, so müssen zum einen der Unterrichtsprozess betrachtet werden, in denen sich dieses Wissen und Können entwickelt hat. Zu den Bedingungen dieses Prozesses gehören die Vorkenntnisse des Schülers, seine individuelle Fähigkeits- und Einstellungsstruktur, der konkrete Verlauf der Unterrichtsabschnitte, das pädagogische Können seines Mathematiklehrers, das Klassen- und Schulklima und viele andere, wobei der Verlauf des Unterrichtsprozesses, d. h. die Inhalte und Unterrichtsmethoden sicher einer der wesentlichen Faktoren ist. Im Ergebnis dieses Prozesses hat sich beim Schüler ein bestimmtes System von Wissens- und Könnenselementen herausgebildet. Die Wahrscheinlichkeit der Qualität seines Wissens und Könnens hängt von den Bedingungen des Unterrichtsprozesses und den anderen Faktoren ab.

Der Test selbst läuft wiederum unter bestimmten Bedingungen ab, zu denen die Qualität des vorhandenen Wissens und Könnens des Schülers im Lösen linearer Gleichungen, seine Vorbereitung auf den Test, die Validität, Reliabilität und Objektivität des Testverfahrens und die psychische Situation des Schülers gehören. Im Ergebnis des Tests entstehen Daten. Es ist das Ziel einer jeden Testentwicklung, das Testverfahren so gut zu machen, dass die Wahrscheinlichkeit der Ausprägungen des untersuchten Merkmals hauptsächlich durch die Qualität des zu messenden Personenmerkmals bestimmt wird.

Man kann zwar (im weiten Sinne des Wortes Bedingung) als eine Bedingung für den Prozess der Erfassung der Daten auch den vorherigen Unterricht bezeichnen, dies würde aber zu einer sehr geschachtelten und vielschichtigen Betrachtung führen. Im Sinne einer strukturierten Betrachtungsweise ist es sinnvoll, bei einem Prozess nur die unmittelbar im Verlauf des Prozesses wirkenden Bedingungen zu betrachten.

Bei der Auswertung statistischer Daten von Leistungserhebungen werden zunächst mit bestimmten statistischen Kenngrößeren oder anderen Darstellungsformen die Merkmale und Besonderheiten des Datensatzes beschrieben. Die Auswertung sollte sich aber nicht darauf beschränken. Vielmehr sollte es um das Suchen nach Zusammenhängen zwischen den Bedingungen des Unterrichtsprozesses, in dem die getesteten Personenmerkmale entstanden sind und den Daten bzw. der Verteilung der Daten gehen. Dazu müssen die Unterrichtsbedingungen möglichst umfassend bekannt sein.

Stern und Hardy (2001) heben im Sinne unsere Auffassung hervor, dass Leistungstests, in denen schulbezogene Kompetenzen erfasst wurden, nur dann sinnvoll interpretierbar sind, wenn Informationen über Lerngelegenheiten einbezogen werden. Wenn zum Beispiel ein Schüler die binomische Formel nicht anwenden kann, weil diese im Mathematikunterricht nie behandelt wurde, habe man andere Schlussfolgerungen zu ziehen, als wenn ein Schüler dies trotz ausgiebiger Übung im Unterricht nicht kann (S. 162).

Ein weiteres Beispiel zur Bekräftigung dieser Auffassungen liefern die vergleichenden Untersuchungen in England und Deutschland von Kaiser-Meßmer und Blum (1993). Sie überprüften das Verständnis des Begriffs Prisma anhand einer Aufgabe, in der die Schüler in gegebenen Schrägbildern die Grundflächen von Prismen markieren sollten. Sie stellten überraschend fest, dass zwei leistungsstarke gymnasiale Klassen weitgehend versagten (nur 9 % bzw. 14 % richtige

Lösungen) während in einer leistungsschwächeren Klasse 69 % richtige Lösungen auftraten. Da die Autoren im vorherige Unterricht Autoren hospitiert hatten, konnten sie eine Erklärung für dieses verblüffende Ergebnis finden: In den leistungsstärkeren Klassen wurden nur wenige ähnliche Aufgaben bearbeitet, während in der leistungsschwächeren Klasse mehrere Aufgaben dieser Art im Unterricht behandelt wurden.

Auch Zech und Wellenreuther (1992) haben bei ihren empirischen Untersuchungen zur Evaluation eines Unterrichtsmaterials darauf geachtet, dass sie wenigstens einmal in 14 Tagen im Unterricht hospitieren konnten, um sich einen ungefähren Eindruck davon zu verschaffen, in welchem Umfang tatsächlich nach ihrer Konzeption unterrichtet wurde. Aufgrund der Knappheit der Mittel haben sie dafür eine geringere Anzahl von Versuchskursen in Kauf genommen. „Eine summative Evaluation ist u. E. nur dann aussagekräftig, wenn die Ausführung der relevanten Bedingungen in gewissem Umfang kontrolliert werden kann." (S. 155) Es ist allerdings zu vermuten, dass nur mit Hospitationen im Unterricht und dann noch in einer so geringen Zahl kaum alle relevanten Bedingungen erfasst werden können.

Leider wird jedoch bei den meisten von uns betrachteten empirischen Untersuchungen und auch den meisten Leistungserhebungen der zuvor abgelaufene Unterricht als eine der wesentlichen Bedingungen des Prozesses der Entwicklung der Personenmerkmale überhaupt nicht erfasst. Zwei typische Beispiele für dieses Desiderat aus der mathematikdidaktischen Forschung sind die Untersuchungen von Padberg und Bienert (2000) und Neumann (2000). In beiden Fällen wurden lediglich Tests mit Schülern durchgeführt, um ihr Verständnis von Begriffen bzw. ihre Rechenfertigkeiten zu überprüfen. Es gibt keinerlei Bemerkungen zum Verlauf des Unterrichts, in dem diese Dispositionen entstanden sind, obwohl es sich bei Padberg und Bienert (2000) um eine Längsschnittstudie handelt. In Auswertung der Daten werden dann aber Folgerungen für den Unterricht abgeleitet.

Ein erstes Merkmal zum Vergleich der Leistungserhebungen soll deshalb die Einbeziehung von **Bedingungen des voran gegangenen Unterrichts** sein. Dabei geht es vor allem um den konkreten Verlauf des Unterrichtsprozesses, also um die Unterrichtszeiten für einzelne Themen bzw. Kompetenzbereiche, die verwendeten Aufgaben, die Art ihrer Bearbeitung u. a. konkreten Merkmale.

Eine wesentliche Voraussetzung zur Untersuchung von Zusammenhängen zwischen Ergebnissen und Bedingungen ist die **Festlegung der Grundgesamtheit und ihrer Elemente** als Bezugssystem der Datenauswertung. Bei jeder Leistungserhebung werden immer Daten von einzelnen Individuen erfasst, wobei wir uns im Folgenden nur auf die Betrachtung von Schülern beschränken wollen. Die *originäre* Grundgesamtheit einer jeden Erhebung besteht zunächst aus der Gesamtheit aller getesteten Schüler. Die Grundgesamtheit kann aber auch eine weit umfassendere Schülerpopulation sein, wenn es sich bei der Erhebung um eine repräsentative Stichprobe aus dieser Population handelt. Bei der Interpretation der Ergebnisse der Erhebung müssen die Bedingungen in der jeweiligen Grundgesamtheit beachtet werden.
Bei einer Klausur besteht die Grundgesamtheit aus allen Schülern einer bestimmten Klasse und der betreffende Lehrer muss bei der Interpretation der Ergebnisse alle Bedingungen beachten, die in seiner Klasse einen Einfluss auf die Entwicklung der überprüften Leistungseigenschaften hatten. Handelt es sich um eine landesweite Erhebung (als Vollerhebung oder als repräsentative Stichprobenerhebung), so kann als Grundgesamtheit die Menge aller Schülern des betreffenden Landes gewählt werden und es können dann bei der Interpretation der Landesergebnisse nur die allgemeinen Bedingungen, die an allen Schulen in allen Klassen des Landes gelten, verwendet werden. Es wird also von den spezifischen Besonderheiten von Schularten, den besonderen Bedingungen an einer Schule bzw. den individuellen Unterschieden von Lehrern abstrahiert.

Die einzelnen Fälle (Schüler) können aber auch zu bestimmten Gruppen zusammengefasst (aggregiert) werden, wobei man sich in der Regel an bildungsorganisatorischen Strukturen orientiert,

also eine Zusammenfassung zu Klassen, Schulen, Bildungsgängen oder Ländern vornimmt. Aber auch eine Einteilung nach Jungen und Mädchen ist üblich. Durch diese Zusammenfassungen entsteht eine Grundgesamtheit mit neuen Elementen, auf die sich die Auswertungen zu beziehen haben. Eine solche Zusammenfassung ist allerdings nur gerechtfertigt, wenn die so entstehenden neuen Objekte der neuen Grundgesamtheit aufgrund der Gemeinsamkeiten in ihren Bedingungen auch als vergleichbar angesehen werden können.
Wird etwa als Grundgesamtheit die Menge aller Schulen eines Landes gewählt, so bedeutet dies, dass die Bedingungen an allen Schulen als vergleichbar angesehen werden und bei der Interpretation einer statistischen Kenngröße in dieser Grundgesamtheit nur das Gemeinsame der Bedingungen betrachtet werden kann.

Die Zusammenfassung von Fällen ist ein kompliziertes und vielschichtiges Problem einer Element-System-Beziehung und kann nicht auf eine rein formale Unterscheidung von Hierarchie-Ebenen (Helmke; Hosenfeld 2004, S. 57) reduziert werden.

Als Beispiel sei auf die kontroverse Mediendebatte zu den herausragenden Ergebnissen der Laborschule Bielefeld und einer weiteren Reformschule bei PISA 2000 verwiesen (Eikenbusch; Leuders 2004), die zu einer Stellungnahme des Max-Planck-Instituts für Bildungsforschung geführt hat, in der die Fehlinterpretation der Daten herausgestellt wurde (Max-Planck-Institut für Bildungsforschung 2002). Aus dieser Stellungnahme ist erkennbar, dass offensichtlich die individuellen Voraussetzungen der Schüler an diesen Reformschulen nicht mit denen an den üblichen Schulen vergleichbar sind. Es ist stark zu vermuten, dass auch andere Bedingungen sich erheblich von denen der üblichen Regelschulen unterscheiden, so dass eine Zusammenfassung der Fälle nach Schulen unter Einbeziehungen dieser Reformschulen nicht gerechtfertigt ist.

Zur Lösung dieses Problems wird bei Schulrückmeldungen in Landeserhebungen oft eine Berücksichtigung von Kontextmerkmalen wie häusliche Lernumwelt, Altershomogenität der Klasse oder Schulgröße vorgenommen und die Schülerdaten werden entsprechend adjustiert (Helmke; Hosenfeld 2004, S. 58).

Das Problem der Vergleichbarkeit von Bedingungen führte bei der internationalen Lesestudie IGLU, an der sich 35 Staaten beteiligt haben, zur Bildung von drei Ländervergleichsgruppen, in die die Länder nach vergleichbaren Kriterien gruppiert wurde, ohne damit alle Länder erfassen zu können (Bos u. a. 2003, S, 9).

Wie bei jedem psychologischen Test ist auch bei einer Leistungserhebung eine entscheidende Ausgangsfrage, die möglichst genaue und konkrete **Bestimmung des zu testenden Personenmerkmals**[2]. Wir fassen ein Personenmerkmal als eine real existierende habituelle psychische Eigenschaft auf, deren konstitutionelle Grundlage ein bestimmtes individuelles System psychischer Zustände und Prozesse ist. Die Zustände und Verlaufseigenschaften der Prozesse sind Ausdruck einer bestimmten Konstellation der neuronalen Strukturen und Zustände im Gehirn des Individuums. Als begriffliche Konstrukte zur Beschreibung der Personenmerkmale werden u. a. die Begriffe Wissen (Kenntnisse), Fähigkeiten, Fertigkeiten, Einstellungen und Können als Komplexe dieser Merkmale verwendet. Mit diesen begrifflichen Konstrukten können sehr begrenzte Systeme psychischer Zustände, wie etwa die Kenntnisse zum Satz über die Innenwinkelsumme im Dreieck, aber auch sehr umfangreiche und vielschichtige Systeme wie etwa das Raumvorstellungsvermögen beschrieben werden. Der Extremfall ist, dass mit einem einzigen Begriff wie dem der „mathematical literacy“ das gesamte System des mathematischen Wissens und Könnens erfasst werden soll. Wir halten es allerdings zur Modellierung der kognitiven Strukturen für sinnvoller, möglichst eng begrenzte Bereiche psychischer Dispositionen als Grundlage für begriffliche Konstrukte zu verwenden.

---

[2] „Personenmerkmal“ und „psychische Disposition“ werden synonym verwendet.

Das zu messende Personenmerkmal wird in der psychologischen Testtheorie auch als latente (verborgene) Variable bezeichnet, vorausgesetzt, das Merkmal lässt sich durch einen einzigen Parameter beschreiben, der quantitativ oder qualitativ skalierbar ist.

Ein grundlegendes Problem bei der Bestimmung der Personenmerkmale ist die verwendete Struktur des mathematischen Wissens und Könnens oder mit anderen Worten die verwendete Taxonomie der Ziele des Mathematikunterrichts. Wir orientieren uns, wie bereits in 1.4.3 dargestellt, an den „Linienführungen“ der Rahmenpläne von Mecklenburg-Vorpommern sowie an dem Modell der Taxonomie der Ziele nach Sill (2002).

Bei einem psychologischen Test und auch bei jeder Leistungserhebung muss die **Anzahl der Dimensionen** festgelegt werden, in denen sich die Erhebung bewegt. Dieselbe Aufgabe kann zur Diagnostik unterschiedlicher Personenmerkmale verwendet werden, wenn die Lösung dieser Aufgabe durch verschiedene Personenmerkmale beeinflusst wird. Die Anzahl der Merkmale, die erfasst werden sollen, wird als Dimension bezeichnet.

Mit der Festlegung der Grundgesamtheit und des verwendeten Testmodells hängt auch das Problem der **Normierung** eines Erhebungsverfahrens zusammen. In der psychologischen Testtheorie wird zwischen einer normorientierten (sozialnormorientierten) und einer kriteriumsorientierten (curriculumbezogenen, lehrzielorientierten) Normierung unterschieden. *Normorientiert* bedeutet, dass die Verhältnisse in der Grundgesamtheit als Maßstab zur Normierung verwendet werden. Diese Verhältnisse werden meist mithilfe einer repräsentativen Stichprobe (Eichstichprobe) ermittelt. Die Bewertung der Leistung eines Schülers bei einem so normierten Erhebungsverfahren ergibt sich dann durch einen Vergleich mit der Leistung aller Schüler der betrachteten Grundgesamtheit.

Normorientiertes Testen ist bei schulischen Leistungsmessungen ein eher konservatives Vorgehen, da der aktuelle Entwicklungsstand als Bezugsmaßstab gewählt wird. Überprüft man die Leistungen der Schüler in gegenwärtig vernachlässigten Bereichen, wie etwa der Stochastik, führt diese Normierungsart zu einer Festschreibung des Zustandes.

Bei einer *kriteriumsorientierten Normierung* wird ein Maßstab von außen vorgegeben. Dies sind im schulischen Bereich die gewünschten bzw. erwarteten Leistungsergebnisse im Unterricht. Die Angabe von Erfüllungsprozenten entspricht in der Regel einer solchen Normierung, da bei der Erstellung des Erhebungsverfahrens die Autoren meist von einer annähernden Normalverteilung im Bereich von 0 bis 100 % der zu erreichende Punktzahl ausgehen.

Durch eine Normierung (Eichung) eines Testes ergibt sich eine Skala mit Vergleichswerten (Normen), mit denen man die Einzelergebnisse eines Testes bewerten kann. Bei einer sozialnormorientierten Eichung ergeben sich die Vergleichswerte aus der Durchführung des Testes in einer Stichprobe. Zur Auswertung der Daten ist die Verwendung eines Testmodells erforderlich, mit dem der Zusammenhang zwischen der Ausprägung des zu messenden Personenmerkmals und den Daten der Tests modelliert wird.

Ein Aspekt zur Beschreibung der Modelle ist die Betrachtung der so genannten Itemfunktion. Ein Item ist die kleinste Beobachtungseinheit eines Tests. Oft werden bei standardisierten Testverfahren dichotome Variable betrachtet, das heißt, es wird nur zwischen richtiger und falscher Beantwortung des Items unterschieden. Eine Itemfunktion (bzw. Itemcharakteristik) beschreibt bei einer dichotomen Variablen die Abhängigkeit der Wahrscheinlichkeit einer richtigen Beantwortung des Items von der Ausprägung des Personenmerkmals. Die Schwierigkeiten (bzw. Leichtigkeit) eines Items ergibt sich aus der Anzahl der richtigen (bzw. falschen Antworten), also dem Erfüllungsprozentsatz einer Aufgabe bei einer Leistungserhebung. Unter der Trennschärfe eines Items versteht man den Anstieg der Itemfunktionen, die größer der Anstieg desto größer die Trennschärfe.

Bei dem heute häufig als Testmodell benutzten Rasch-Modell ist die vorausgesetzte Itemfunktion eine logistische Funktion. Das Raschmodell ist in seiner Urform ein eindimensionales Modell, d. h. es

wird nur ein Personenmerkmal (θ) betrachtet. Jedem Item wird ein (empirisch ermittelter) Schwierigkeitsparameter (σ) zugeordnet. Die Funktion für ein Item mit der Schwierigkeit σ hat dann die Form $f(\theta) = \frac{1}{1+e^{(\sigma-\theta)}}$.

Bei internationalen Leistungserhebungen und bei Erhebungen, die sich an diesen orientieren, werden oft Aufgaben mit Mehrfachwahlantworten (Multiple-Choice-Aufgaben) eingesetzt. Deshalb soll auch das Merkmal **Antwortformat** der verwendeten Aufgaben betrachtet werden.

Bei den üblichen Aufgaben in Schullehrbüchern und entsprechend auch in Klausuren und Prüfungsarbeiten muss der Schüler in der Regel einen Lösungsweg und die Lösung selbst angeben. Dies wird von Psychologen als *freies Antwortformat* bezeichnet. Leistungserhebungen mit freien Antwortformaten entsprechen also den Aufgabenstellungen, die der Schüler aus dem Unterricht gewohnt ist. Sie sind in der Regel nicht durch Erraten der Antwort lösbar, sondern zwingen den Schüler, die erforderlichen geistigen Handlungen und schriftlichen Tätigkeiten, deren Ausführbarkeit mit der Aufgabe überprüft werden soll, auch tatsächlich vorzunehmen.

Bei einem *gebundenen Antwortformat* wird dem Schüler zusammen mit der Aufgabenstellung eine Anzahl von Antwortmöglichkeiten vorgegeben, aus denen er in der Regel eine oder manchmal auch mehrere auszuwählen hat. Bei dieser Art der Aufgabenstellung besteht die einzige schriftliche Tätigkeit des Schülers im Markieren von Antwortmöglichkeiten. Dies kann ihn zu oberflächlichen Entscheidungen oder sogar zum Raten veranlassen, worauf Meyerhöfer (2004a) in seiner Kritik an PISA hinweist. Wie die Untersuchungen von Woschek (2004) mit TIMSS-Aufgaben zeigen, raten dabei allerdings nur etwa 5 bis 10 % der Schüler. Die vorzunehmenden geistigen Operationen entsprechen aber in der Regel nicht denen bei der gleichen Aufgabe mit freiem Antwortformat.

Der Schüler kann in einigen Fällen sehr unterschiedlich vorgehen. Soll etwa eine einfache Gleichung gelöst werden und sind mehrere Lösungen vorgegeben, so kann der Schüler die Gleichung im Kopf lösen und seine Lösung mit den Antworten vergleichen oder er überprüft, ob die vorgegebenen Zahlen die Gleichung erfüllen. Wenn er Kenntnisse zur maximalen Anzahl der Lösungen der Gleichung besitzt, hat er dann eine unterschiedliche Anzahl von Berechnungen durchzuführen, je nachdem mit welcher Zahl er anfängt (vgl. Beispielaufgabe 4, Helmke; Jäger 2002, S. 61). Die vorgegebenen Auswahlantworten sind nur in seltenen Fällen exhaustiv, d. h. schöpfen alle Antwortmöglichkeiten aus. Mit den falschen vorgegebenen Antwortmöglichkeiten (Distraktoren) kann man zwar einige typische Fehler erfassen, aber im tatsächlichen Lösungsverhalten von Schülern treten meist weit mehr Varianten fehlerhafter Antworten auf, wie auch unsere Erhebungen zeigen.

Der einzige Vorteil der gebundenen Antwortformate ist die Auswertungsökonomie. Es ist eine sehr schnelle und oft auch elektronische Auswertung der Eintragungen der Schüler möglich, was besonders bei großen Populationen vorteilhaft ist.

Ein weiteres Merkmal einer Leistungserhebung ist die **Art der speziellen Vorbereitung** der Lernenden auf diese Erhebung. Damit im Zusammenhang steht der Geheimhaltungs- bzw. **Bekanntheitsgrad der Anforderungen** für Lernende und Lehrende. Die konkreten Anforderungen von Leistungserhebungen sowie ihr Termin können den Lernenden weitgehend bekannt sein. Dies ermöglicht eine gründliche und spezielle Vorbereitung der Lernenden.
Die Leistungserhebung kann aber auch völlig unvorbereitet für den Lernenden mit vorher geheim gehaltenen Aufgaben erfolgen. Bei einer unvorbereiteten Leistungserhebung kann praktisch das Leistungsniveau festgestellt werden, was dauerhaft und präsent angeeignet wurde und damit jederzeit abrufbar ist. Dies kann sich erheblich von dem unterscheiden, was nach einer zielgerichteten Vorbereitung z. B. auf eine Klausur an Leistungen erbracht werden kann. Viele Anforderungen im Alltag, in anderen Unterrichtsfächern oder auch in der Berufsausbildung bzw. im Studium entsprechen denen einer unvorbereiteten Leistungserhebung. Es gibt aber auch zahlreiche Situationen in der späteren Ausbildung, in denen in möglichst kurzer Zeit eine selbständige oder angeleitete Reaktivierung von Wissen und Können erfolgen muss. Deshalb wäre es durchaus

angebracht, eine ähnliche Situation auch bei einer Leistungserhebung zu organisieren. Die Prüflinge können z. B. vor dem Lösen der Aufgaben ein schriftliches Material erhalten, das sie in einer vorgegebenen Zeit durcharbeiten können und das die fachlichen Grundlagen für die Aufgaben erhält. Mit der Leistungserhebung wird dann ermittelt, wie schnell nicht mehr bewusstes Wissen und Können reaktiviert werden kann, was auch davon abhängt, wie gut es bei seiner Vermittlung strukturiert und in das vorhandene Wissen und Können integriert wurde.

Es sollen ebenfalls die **Durchführungsbedingungen** als ein Merkmal der Leistungserhebung betrachtet werden. Dabei geht es z. B. um die Organisation, den Zeitumfang, der rechtliche Rahmen oder den Turnus der Erhebung.

Eine generelle Anforderung an jede Evaluation besteht nach Lind (2003) darin, dass die Bedürfnisse und Interessen der Prüfenden und der Prüflinge bei der Vorbereitung, Durchführung und Auswertung einer Erhebung berücksichtigt werden müssen. Dies vor allem auch deshalb erforderlich, damit die Ergebnisse der Erhebung von allen akzeptiert und die sich daraus ergebenden Konsequenzen von allen mitgetragen werden. Dies ist sowohl eine Konsequenz moralisch-demokratischer Forderungen, aber nach Wottawa; Thierau (1998, S. 69) ist die „repräsentative Beteiligung aller Betroffenen an der Planung und Durchführung der Erhebungen“ eine grundlegende Anforderung an jede Evaluationsstudie. Deshalb soll die **Partizipation** an der Erhebung beteiligter Personen und Institutionen als ein Merkmal der Erhebung betrachtet werden.

Das in einer Leistungserhebung verwendetet Erhebungsverfahren kann zur Erfassung der Ausprägungen der Personenmerkmale mehr oder weniger geeignet sein. Die Güte eines Verfahrens kann durch die Gütekriterien für psychologische Tests beschreiben werden (vgl. Rost 2004, S. 33 ff.). Diese Gütekriterien erfordern bestimmte Methoden zur Entwicklung der Erhebungsverfahren wie etwa die Pilotierung der Aufgaben. Deshalb sollen die **Entwicklung und Güte der Testverfahren**[3] weitere Vergleichsmerkmale sein.

Es sollen weiterhin die **Methoden zur Auswertung der Daten** verglichen werden. Dies beginnt bei der Art der Erfassung der Daten. Sie reicht von der Erfassung und Auswertung aller originalen Schülerlösungen, wie wir es bei den Vergleichsarbeiten angestrebt haben bis zur lediglich dichotomen Erfassung der Ergebnisse und der reinen Angabe von Erfüllungsquoten. Die Methoden der Auswertung hängen u. a. von den Aufgabenformaten und dem verwendeten Testmodell ab. Ein Testmodell, wie das Rasch-Modell, ist ein formales mathematisches Modell, das sich aus theoretischen Annahmen über Art der getesteten Personenmerkmale und aus den Eigenschaften der vorliegenden Daten ergibt. Die Auswertung der Daten sollte weiterhin transparent, d. h für alle verständlich und nachvollziehbar sein (Lind 2003).

Zur Auswertung der Daten gehört ebenfalls, ob die Leistungen der Schüler bewertet bzw. zensiert werden.

Bei allen Funktionen wird auch die mögliche **Rolle der Leistungserhebungen für die didaktische Forschung** diskutiert. Eine zentrale Aufgabe der Didaktik als einer empirischen Wissenschaft ist die Ermittlung von Gesetzen des Lernens und Lehrens in einem Fachgebiet. Leistungserhebungen können wesentliche Bestandteile entsprechender empirischer Untersuchungen sein.

2.2. Funktionen von Leistungserhebungen

2.2.1. Grundlagen der Funktionsbetrachtungen

Weinert (2001a) betont, dass beim Einsatz von Verfahren zur Leistungsmessung stets die beabsichtigte Nutzung der Daten beachtet werden muss. Er unterscheidet zu diesem Zweck

[3] „Testverfahren“ wird hier synonym zum Wort „Erhebungsverfahren“ verwendet und hat nicht alle Bedeutungen des Begriffes „Test“, den wir im Sinne der psychologische Testtheorie (z. B. Rost 2004, S. 17) verwenden.

verschiedenen Ebenen der Leistungsmessung, die mit den Ebenen des Schulsystems in Verbindung stehen und auf denen die Messungen jeweils spezielle Ziele haben (S. 28 ff.). Er stellt fest, dass es auf der Ebene eines Staates oder Bundeslandes um bildungspolitische Entscheidungen geht. Hier schaffen Vergleichsuntersuchungen ein Orientierungswissen, das zur Bildung von Hypothesen beitragen kann, ohne dass die Ergebnisse solcher Leistungserhebungen geeignet sind, bildungspolitische Entscheidung direkt zu fundieren oder zu steuern.

Als zweite Ebene betrachtet er landesweite oder regionale Orientierungsstudien. Ziel dieser Arbeiten sei es, allen am Bildungsgeschehen beteiligten Informationen über den Entwicklungsstand bestimmte Kompetenzen zu geben.

Die dritte Ebene ist nach seiner Auffassung die schulische und unterrichtliche Qualitätsentwicklung, die durch angemessene Rückmeldung an Schulen oder Klassen über relevante Ergebnisse landesweiter Leistungserhebungen aber auch durch viele kleine möglichst häufig durchgeführte Arbeiten auf lokaler schulischer Ebene erfolgen könnte.

Der Strukturierungsvorschlag von Weinert für Ziele von Leistungserhebungen wie auch Vorschläge andere Bildungsforscher (Helmke; Hosenfeld 2003) sind vor allem an äußeren Merkmalen orientiert. Es werden weiterhin gleichzeitig verschiedene Aspekte wie Entscheidungsfindung, Informationsvermittlung und Entwicklungsprozesse betrachtet.

Um einen Vergleich der Ziele von Leistungserhebungen vorzunehmen, wird von uns ein theoretischer Ansatz verwendet, der sich bereits bei der Analyse von unterschiedlichen Zielformulierungen für die mathematische Allgemeinbildung sowie zur Bestimmung von Zielen mathematischen Wahlunterrichts bewährt hat (Sill 1997). Es hat sich dabei als konstruktiv erwiesen, Funktionen und Ziele pädagogischer Prozesse zu unterscheiden. Als pädagogische Prozesse werden alle Vorgänge bezeichnet, in denen durch „Erzieher“ in zielgerichteter Weise Personenmerkmale von „Zöglingen“ verändert werden sollen. Wir beschränken uns auf pädagogische Prozesse im Rahmen des Bildungssystems und dabei vor allem auf den Klassenunterricht.

Unter eine Funktion verstehen wir die Rolle eines Systemelements. Zur Angabe einer Funktion ist deshalb immer die Angabe eines Bezugssystems erforderlich. Bezugssysteme können wieder pädagogische aber auch andere Prozesse sein.

Während Ziele reine Auflistungen von Absichten sind, die unter sehr unterschiedlichen Aspekten entstanden sein können, geht es bei einer Funktion um eine beabsichtigte Wirkung im Rahmen eines Systems von Prozessen. Ein bestimmtes Ziel kann durchaus unterschiedlichen Funktionen zugeordnet werden. Formulierungen von Funktionen und Zielen sind sprachlich sehr ähnlich. Jede Funktionsangabe kann auch als Zielformulierung angesehen werden. Zielformulierungen können auch das Wort Funktion enthalten. Der entscheidende Unterschied ist die systemische Sichtweise. In der Literatur wird oft nicht zwischen Funktionen, Zielen oder Aufgaben unterschieden und die Wörter werden oft synonym verwendet.

Mit den von uns vorgeschlagenen Betrachtungen zu Funktionen von Leistungserhebungen lässt sich die Fülle der möglichen Aussagen zu den Zielen von Erhebungen in geeigneter Weise kohärent strukturieren. Aus der Funktion einer Leistungserhebung ergeben sich ihre speziellen Ziele sowie die Ausprägung der zuvor diskutierten Merkmale. Eine Beurteilung dieser Merkmale ohne Beachtung der intendierten Funktion ist u. E. nicht sinnvoll.

In den folgenden Betrachtungen beschränken wir uns (zumindest im eigenen Denken) stets auf Leistungserhebungen zum Mathematikunterricht bzw. zur mathematischen Bildung. Vieles lässt sich analog sicher auch auf andere Bildungsbereiche übertragen.

Als *Bezugssysteme* für die Funktionen von Leistungserhebungen werden im Sinne einer dynamischen und entwicklungspsychologisch orientierten Sicht *Entwicklungsprozesse* gewählt.

Wir unterscheiden *drei verschiedene Gruppen* von Entwicklungsprozessen:

- die Entwicklung des mathematischen Wissens und Könnens von Lernenden
- die Entwicklung des beruflichen Könnens von Lehrern sowie die Entwicklung von Fachkollegien
- die Entwicklung von Rahmenbedingungen von Schule und Unterricht wie die Struktur des Bildungssystems, die Gestaltung von Curricula oder die Entwicklung von Testverfahren

Die Funktionen von Leistungserhebungen in diesen Bezugssystemen sollen im Folgenden anhand der vorher diskutierten Merkmale beschrieben werden. Dabei werden zu Illustration bereits Beispiele von Leistungserhebungen angeführt. Eine ausführlichere Diskussion der im 1. Kapitel beschriebenen Leistungserhebungen erfolgt erst im Abschnitt 2.3.

Eine bestimmte Form der Leistungserhebung kann mehrere Funktionen haben, d. h. in verschiedene Bezugssysteme eingeordnet werden. Die bei jeder Funktionsart jeweils gewählten Beispiele haben hauptsächlich die betreffende Funktion. Auf weitere Funktionen wird nur kurz eingegangen.

2.2.2. Funktionen in der Entwicklung des mathematischen Wissens und Könnens von Lernenden

Es können in diesem Bezugssystem zwei Funktionen unterschieden werden:

- ***Funktion der Leistungsbewertung***
- ***Funktion der individuellen Diagnostik***

Leistungserhebungen mit diesen Funktionen haben eine Reihe gemeinsamer aber auch unterschiedlicher Merkmale. Deshalb sollen die Funktionen getrennt betrachtet werden.

Funktion der Leistungsbewertung

Bei dieser Funktion geht es um das Feststellen und Bewerten des persönlichen Leistungsstandes von Lernenden. Typische Formen von Leistungserhebungen, die vor allem diese Funktion haben, sind Klausuren, Abschlussprüfungen oder Einstellungstests. Diese drei Formen werden in den folgenden Betrachtungen als Paradigmen gewählt. Weitere Formen sind tägliche Übungen, Kurzkontrollen oder Klassenarbeiten.

Die **Bedingungen**, unter denen die überprüften Personenmerkmale entstanden sind, also vor allem der Verlauf des vorherigen Unterrichts aber auch die Besonderheiten der Schüler der Klasse werden bei den Analysen der Ergebnisse dieser Leistungserhebungen durch den unterrichtenden Lehrer in der Regel beachtet und führen ihn zu entsprechenden Konsequenzen für seinen weiteren Unterricht. Die Praxis der Länder mit regelmäßigen zentralen Abschlussprüfungen zeigt z. B., dass durch diese Prüfungen ein erheblicher Einfluss auf den Mathematikunterricht in den oberen Klassen besteht. Die Lehrer orientieren sich bei der Gestaltung ihres Unterrichts in starkem Maße an den Anforderungen, die sich aus den bisherigen Prüfungsaufgaben ergeben und an den dabei durch sie erzielten Ergebnissen.

Die originäre **Grundgesamtheit** bei einer Klausur sind die Schüler der jeweiligen Klasse, bei einer zentralen Abschlussprüfung sind es die Schüler der Abschlussklassen des Landes und bei einem Einstellungstest die Menge der möglichen Bewerber.

Die **Bestimmung der zu testenden Personenmerkmale** führt in der Regel zu sehr konkreten und speziellen Komponenten. Bei einer Klausur geht es primär um die speziellen mathematischen Anforderungen des gerade behandelten Stoffgebietes. Allgemeine Dispositionen wie Abstraktionsvermögen oder Argumentationsfähigkeit werden wohl kaum durch den Lehrer als Testgegenstand ausgewählt werden.

Auch wenn es in Klausuren und Abschlussprüfungen am Ende nur um eine Zensur geht, mit der die Gesamtleistung des Schülers bewertet werden soll, sind diese Erhebungen stets **mehrdimensional**. So wird z. B. ein Lehrer mit einer Klausur zu linearen Gleichungen mindestens die folgenden Merkmale testen wollen: Kenntnisse zu den Umformungsregeln für Gleichungen, Können im Umformen von Termen, Können im Anwenden der Regeln zum formalen Lösen von Gleichungen,

Einstellungen und Verfahrenskenntnisse zur Kontrolle der Lösungen von Gleichungen sowie Können im Lösen von Sachaufgaben. Auf Grund der beschränkten Zeit ist es nur möglich, eine Auswahl von Aufgaben zu den einzelnen Dimensionen vorzunehmen. Mit einer Aufgabe wie etwa einer Sachaufgabe können Aussagen zu mehreren Dimensionen erhalten werden.

Bei Abschlussprüfungen ist die Anzahl der Dimensionen noch weit größer und die Beschränkung der Itemzahl erlaubt noch weit weniger Informationen zu den einzelnen Merkmalen.

Die Leistungserhebungen sind stets **normiert**, da es letztlich um eine Bewertung des Prüflings geht. Die Normierung von Klausuren und Abschlussprüfungen erfolgt scheinbar kriteriumsorientiert (lehrzielorientiert), da beim Schreiben von Klausuren oder Abschlussprüfungen die Leistung an den Erfüllungsprozenten gemessen wird und es in der Regel festgelegte Zuordnungen von Prozentbereichen und Zensuren gibt. Allerdings gibt es keine verbindlichen festgelegten Normen, so dass Lehrer und Mitglieder von Aufgabenkommissionen bei jeder Arbeit selbst einen solchen Bezugsrahmen festlegen müssen. Dabei orientieren sie sich explizit oder implizit an der aktuellen Leistungsverteilung in der betreffenden Grundgesamtheit. Damit erfolgt die Normierung de facto normorientiert.

Dies bestätigen Untersuchungen von Ingenkamp (1997, S. 110) und auch Untersuchungsergebnisse im Rahmen der PISA-Tests. Bei der Analyse der erfassten Mathematiknoten und ihrem Vergleich mit den im PISA-Test 2000 ermittelten Mathematikleistungen zeigte sich, dass 37 % der Varianz der Mathematikleistung auf Unterschiede in den Bildungsgängen entfallen, während bei den Mathematiknoten 90 % der Unterschiede in den Schulen liegen und der Bildungsgang fast gar nicht differenziert. Dies bedeutet, dass die Lehrer bei ihren Bewertungsmaßstäben als Referenzrahmen die eigene Klasse verwenden, so dass Noten aus verschiedenen Klassen, Schulen oder Bildungsgängen nicht miteinander vergleichbar sind.

Aus Sicht der Schüler ist ein sozialnormorientiertes Vorgehen zur Bewertung ihrer Leistungen durchaus gerecht, da sie nicht für die Mängel in der Unterrichtsgestaltung oder die Zusammensetzung ihrer Klasse verantwortlich gemacht werden können.

Das **Antwortformat** der Aufgaben in Klausuren und Abschlussprüfungen entspricht dem üblichen Format in Schullehrbüchern und im Unterricht. Es ist nicht zu verantworten, die Schüler in solchen Situationen mir ungewohnten Fragestellungen zu konfrontieren. Das freie Antwortformat ermöglicht zudem dem Lehrer eine viel tiefgründigere Auswertung der Schülerleistungen.
Bei Einstellungstests sind gebundene Antwortformate sinnvoll, da es nur um eine vergleichende Einschätzung und nicht um eine Zensierung oder gar Fehleranalyse geht und eine schnelle Auswertung der Tests wünschenswert ist.

Die Anforderungen der bisherigen Leistungserhebungen sind bei Klausuren in Form der bisher im Unterricht bearbeiteten Aufgaben und bei Abschlussprüfungen durch die Veröffentlichung bisherigen Prüfungsaufgaben weitgehend bekannt. Der Prüfling kann sich zielgerichtet und intensiv auf die Leistungserhebung **vorbereiten**.

Bei Einstellungstests werden zwar die Aufgaben nicht veröffentlicht, da in der Regel dieselben Tests verwendet werden, aber die Kandidaten haben genügend Möglichkeiten Informationen zu erhalten (z. B. analoge Trainingsaufgaben), um sich intensiv und zielgerichtet auf den Test vorzubereiten.

Durch eine intensive zielgerichtete Vorbereitung auf die Leistungserhebung, die von allen Beteiligten gewollt bzw. erwünscht ist, ergeben sich meist zwei Effekte. Solche Vorbereitungsphasen insbesondere bei Abschlussprüfungen bewirken eine erhebliche Festigung des Wissens und Könnens und oft sogar eine Neustrukturierung und Integration der Kenntnisse, die zu neuen Einsichten und Dispositionen führen. Solche Phasen sind also für den Lernprozess sehr günstig, wenn nicht sogar notwendig. Auf der anderen Seite entspricht das aktuelle Leistungsvermögen in einer solchen Prüfung nicht dem tatsächlichen dauerhaften Vermögen. Die Ausmaße dieses Prüfungseffektes, die sich in einem oft weit über dem tatsächlichen Vermögen konstatierten Niveau zeigen, lassen sich

schwer abschätzen. Es müssten dazu empirische Untersuchungen zum Leistungsvermögen vor der Vorbereitungsphase und zur Dauerhaftigkeit nach der Prüfung erfolgen.

Ein Beispiel für diesen Effekt ist das beständige Staunen von Hochschullehrern im Fach Mathematik über den Widerspruch zwischen den völlig unzureichenden Kenntnissen von Abiturienten auf dem Gebiet der Analysis und ihren guten Abiturleistungen auf diesem Gebiet.

Es ist aber auch möglich, dass die Leistungserhebung unangekündigt erfolgt, wie dies etwa bei Kurzkontrollen und täglichen Übungen, die nicht zum laufenden Stoff durchgeführt werden, vorkommen kann. In diesen Fällen sollten aber den Schülern die Anforderungen im Prinzip bekannt sein, so dass sie sich in selbständiger kontinuierlicher Arbeit darauf vorbereiten können.

Zu den **Durchführungsbedingungen** von Klausuren und Abschlussprüfungen gehört, dass sie in einem bestimmten rechtlich abgesicherten Rahmen ablaufen, der gewährleistet, dass nachvollziehbare und gerechte Bedingungen für die Leistungserhebung vorhanden sind.

Die Leistungserhebung selbst stellt allerdings in der Regel eine erhebliche Stresssituation für den Prüfling dar, wodurch das aktuelle Ergebnis beeinflusst werden kann.

Die Häufigkeit und der Turnus dieser Leistungserhebungen sind durch die rechtlichen und sonstigen Rahmenbedingungen festgelegt. Abschlussprüfungen werden z. B. jedes Jahr geschrieben, so dass sich automatisch bestimmte Verfahrens- und Verhaltensweisen der Beteiligten einstellen und auch die Art der Anforderungen eine gewisse Kontinuität hat.

Die **partizipatorischen Anforderungen** sind in der Regel erfüllt, da Prüfer und Prüflinge selbstbestimmt an den Erhebungen teilnehmen und ihre Interessen im Rahmen der rechtlichen Bedingungen beachtet werden.

Nach der Ansicht von Psychologen genügen Leistungserhebungen in der Schule den **Gütekriterien** von psychodiagnostischen Tests oft kaum. Ingenkamp (1997) führt als Beleg zahlreiche Untersuchungen zu schriftlichen Prüfungsarbeiten an (S. 106 ff.). Das Gemeinsame dieser Untersuchungen ist, dass die Ursache für die ungenügende Erfüllung der Gütekriterien stets in der Bewertung der Schülerarbeiten durch die Lehrer gesehen wurde. So haben z. B. verschiedene Lehrer dieselbe Arbeit sehr unterschiedlich bewertet. Das gleiche Phänomen zeigte sich, als Arbeiten von denselben Lehren zweimal in mehrwöchigem Abstand zu bewerten waren. Die Ursachen für diese Unterschiede wurden kaum betrachtet. Sie lagen vermutlich nicht in unterschiedlichen Auffassungen über die Richtigkeit der Lösungen, sondern, wie bei einer der Untersuchungen auch festgestellt wurde, in verschiedenen Auffassungen über die Art der Darstellung der Lösungen.
Die unzureichende Interpretationsobjektivität ergibt sich nach Ingenkamp (1997) aus der Verwendung eines klasseninternen Bezugssystems für die Normierung der Leistungen, so dass gleiche Noten in verschiedenen Klassen sehr unterschiedliche Leistungsniveaus repräsentieren können. Da die Objektivität der schriftlichen Schülerarbeiten nicht gegeben ist, sind Betrachtungen über die Reliabilität und Validität dieser Leistungserhebungen nicht mehr erforderlich, da diese erst bei vorliegender Objektivität diskutabel sind. Sie werden von Ingenkamp auch nicht angestellt, obwohl es leicht möglich wäre, von Lehrern ausgearbeitete Schülerarbeiten auf Reliabilität zu untersuchen. Dazu müsste nur die Bewertung durch andere Personen nach einheitlichen Kriterien erfolgen.

Diese Kritik von Psychologen an den diagnostischen Fähigkeiten der Lehrer reduziert sich also auf das ohnehin bestehende Defizit einer fehlenden schulinternen und auch schulübergreifenden Verständigung der Lehrer über die Maßstäbe der Bewertung von Schülerleistungen.

Es ist allerdings anzunehmen, dass die Daten aus Klausuren und Abschlussprüfungen auch aus anderen Gründen kaum zuverlässig und gültig sind. Wie schon erwähnt, werden durch eine intensive Vorbereitung auf die Leistungserhebung oft nur kurzzeitig die Leistungsdispositionen auf den durch diese Erhebungen ermittelten Stand gebracht. So war es z. B. in den zentralen Abschlussprüfungen

in der DDR in der Klasse 10 üblich, dass die erste Aufgabe eine recht anspruchsvolle Sachaufgabe aus dem Bereich der Prozentrechnung war. Obwohl in der Regel Erfüllungsquoten von über 80 % erreicht wurden, waren aus der anschließenden Berufsausbildung immer massive Klagen zu hören, dass die Schüler einfachste Prozentaufgaben nicht lösen können.

Eine spezielle Vorbereitung auf solche Leistungserhebungen ist zum einen nicht zu verhindern und zum anderen im Interesse der geprüften Person durchaus zu verantworten. Durch geeignete Maßnahmen (vgl. Kap. 6) kann man die Güte der Erhebungen zwar verbessern, aber die Gütekriterien psychologischer Testverfahren werden nie in hoher Qualität zu erfüllen sein.

Bei der **Auswertung** der Ergebnisse von Klausuren und Abschlussprüfungen werden durch die Lehrer meist nur die Punktzahlen für die einzelnen Aufgaben erfasst, aus denen sich dann die prozentuale Erfüllung der Gesamtpunktzahl und die entsprechende Zensur ergeben. In seltenen Fällen erfolgt eine meist qualitativ angelegte Fehleranalyse. Die Auswertungsmethoden sind für Schüler und Lehrer aber transparent.

Für die **didaktische Forschung** stellen Leistungserhebungen mit dieser Funktion keinen geeigneten Gegenstand dar. Ein Hauptproblem sehen wir darin, dass es kaum möglich ist, den Einfluss der gezielten Vorbereitung der Schüler auf diese Erhebungen und die Auswirkungen der Stresssituation zu erfassen. Bei Klausuren und Abschlussprüfungen müssen die Lehrer bzw. Mitglieder der Aufgabenkommissionen zudem in eigener Verantwortung unter Abwägung zahlreicher Aspekte die Aufgaben zusammenstellen, so dass Forscher nur einen geringen Einfluss nehmen können. Wissenschaftliche Experimente sind mit Blick auf die Schüler nicht zu verantworten.

Klausuren und Abschlussprüfungen können auch eine Rolle im nächsten Bezugssystem, der Entwicklung des beruflichen Könnens von Lehren und Fachkollegien spielen. So war die Auswertung der zentralen Abschlussprüfungen in der DDR ein fester Bestandteil der Aktivitäten der Abteilung Mathematik der Akademie der Pädagogischen Wissenschaften, der Fachberater und Fachkonferenzen und auch heute werden in Ländern mit zentralen Abschlussprüfungen jährliche Auswertungen vorgenommen und Lehrer zur Diskussion darüber angeregt (Herwig; Böttner 2001). Auf Potenzen und Grenzen dieser Funktion von Abschlussprüfungen wird im Kap. 6 eingegangen.

Funktion der individuellen Diagnostik

Bei dieser Funktion geht es um das Feststellen von behebbaren Defiziten in den Leistungseigenschaften von Lernenden. Diese Funktion wird vor allem den landesweiten Vergleichsarbeiten zugewiesenen.

Da es um die Diagnose von Besonderheiten des Entwicklungsstandes eines Schülers oder einer Schülergruppe geht, müssen bei der Untersuchung der **Bedingungen des voran gegangenen Unterrichts** vor allem die Bedingungen der individuellen Lernentwicklung der betreffenden Schüler betrachtet werden. Dazu gehören die individuellen Lernvoraussetzungen wie etwa die Entwicklung der allgemein geistigen Fähigkeiten, die Lerneinstellungen aber auch soziale Faktoren wie das häusliche Umfeld. In gemeinsamen Gesprächen der Lehrkräfte mit den Eltern und den Schülern sollte dann festgestellt werden, ob die Defizite behebbar sind und wie dies am besten durch individuelle Fördermaßnahmen erfolgen könnte. Wenn es sich um größere Gruppen von Schülern oder gar eine ganze Klasse handelt, sind Maßnahmen im Klassenverband angebracht.

Bei der originären **Grundgesamtheit** kann es sich je nach Art der Leistungserhebung um die Schüler einer Klasse oder auch die Schüler eines ganzen Landes handeln. Da es nicht um die Feststellung des insgesamt erreichten Entwicklungsstandes geht, ist eine Zusammenfassung einzelner Schülergruppen (z. B. Bildungsgänge) nicht erforderlich.

Bei den zu testenden **Personenmerkmalen** sollte es sich um möglichst eng begrenzte Systeme psychischer Dispositionen handeln. Mit allgemeinen Feststellungen, wie etwa der Aussage, dass sich

der Gesamtleistungsstand eines Schülers unterhalb eines erforderlichen Niveaus befindet, sind kaum zielgerichtete und motivierbare Schritte zur Behebung der Defizite möglich.

Um möglichst genaue Aussagen über ein zu diagnostizierendes Personenmerkmal zu erhalten, ist eine entsprechende Anzahl von Items erforderlich. In Abhängigkeit von der möglichen Länge der Leistungserhebung kann nur eine beschränkte **Anzahl von Dimensionen** betrachtet werden. Denkbar wäre auch ein mehrschrittiges Verfahren, indem zunächst mit einer größeren Anzahl von Dimensionen gearbeitet wird, um erste Vermutungen über mögliche Defizite zu erhalten. In weiteren Schritten könnten dann die betreffenden Personenmerkmale, bei denen die Defizite vermutet werden, mit spezielleren Testverfahren näher untersucht werden.

Um überhaupt von Defiziten sprechen zu können, ist eine **Normierung** der Leistungserhebung erforderlich. Bei kleineren Grundgesamtheiten wie etwa einer Schulklasse kann es sich nicht um eine normorientierte Normierung handeln, da sich dadurch falsche Maßstäbe ergeben können. Ist die Grundgesamtheit z. B. die Menge aller Schüler einer Jahrgangsstufe eines ganzen Landes, so kann die ermittelte Leistungsverteilung eine erste Grundlage für Defizitbeschreibungen sein. Ziel der Entwicklung von Leistungserhebungen für diese Funktion sollte es aber sein, dass diese Verfahren kriteriumsorientiert normiert sind. Insbesondere sollte eine Mindestgrenze der Leistungsdispositionen angegeben werden.

Bei den **Antwortformaten** kann es sich auch um gebundene Formate handeln, wenn dies aus testdiagnostischer Sicht zu ausreichenden Erkenntnissen führt.

Um den tatsächlichen Entwicklungsstand der Personenmerkmale real einschätzen zu können, sollten diese Leistungserhebungen **unvorbereitet** erfolgen.

Die **Häufigkeit,** Inhalt und Umfang dieser Leistungserhebungen sollten gemeinsam durch Lehrer und Schüler vereinbart werden. Ohne eine bewusste **Partizipation** der Schüler können weder unverfälschte Ergebnisse erzielt noch im Anschluss an die Erhebung entsprechende Maßnahmen gemeinsam vereinbart werden.

Die eingesetzten Erhebungsverfahren sollten den **Gütekriterien** psychologische Tests möglichst weitgehend entsprechen. Dies bedeutet gegenwärtig, dass diese Verfahren erst noch zu entwickeln sind (vgl. Funktion 3c).

Die **Auswertung der Daten** sollte für Lehrer und Schüler verständlich sein und kann sich durchaus auf Erfüllungsquoten beschränken, wenn entsprechende kriteriale Normen vorhanden sind. Die Schülerleistungen sollten prinzipiell nicht zensiert werden, um damit verbundene Vorbereitungseffekte und ungerechte Bewertungen von Schülern zu vermeiden.

Solche Leistungserhebungen sind ein Bestandteil **didaktischer Forschungen**, da die differenzierte Arbeit mit Schülern zum Gegenstand der Didaktik gehört. Weiterhin sollten Didaktiker unbedingt an der Entwicklung der Testverfahren beteiligt werden.

2.2.3. Funktion der Entwicklung des beruflichen Könnens von Lehrern und Fachkollegien

Leistungserhebungen in diesem Bezugssystem dienen vor allem der eigenen Überprüfung der unterrichtlichen Bemühungen von Lehrern bzw. Fachkollegien. Unter Fachkollegien werden Fachschaften und Fachkonferenzen einer Schule verstanden.

Geeignete Formen von Leistungserhebungen mit dieser Funktion sind tägliche Übungen, Kurzkontrollen, ein durch einen einzelnen Lehrer sowie auch gemeinsam in der Fachschaft einer Schule oder kooperierender Schulen erarbeitete Testarbeiten oder ein bereitgestellter normierter Test. Diese Erhebungen dienen nicht primär der Bewertung der Schülerleistungen. Ob eine Bewertung trotzdem erfolgen kann, muss in jedem Einzelfall überprüft werden.

Die hier betrachteten Vorgänge werden meist als Schul- und Unterrichtsentwicklung bezeichnet. Die Entwicklung von Schule und Unterricht heißt aber vor allem Entwicklung bzw. Veränderung des

Handelns von Lehrern und Fachkollegien. Dabei können Leistungserhebungen eine wichtige Rolle bei der Initiierung und Evaluation solcher Prozesse spielen.

Die ermittelten Ergebnisse werden schon vom Anliegen her durch die Lehrer und Fachkollegien mit den **Bedingungen** der abgelaufenen Unterrichtsprozesse in Beziehung gesetzt. Der konkrete Verlauf des Unterrichts, die Rahmenbedingungen der Schule sowie die spezifischen Leistungsvoraussetzungen in den jeweiligen Klassen können durch die unterrichtenden Lehrer am besten eingeschätzt werden. Auf der Basis der Ergebnisse der Leistungserhebungen kann dann eine entsprechende Diskussion und Ursachenforschung in der Fachschaft der Schule erfolgen und es können Schlussfolgerungen für ein verändertes Handeln gezogen werden.

Die **Grundgesamtheit** bei einer solchen Erhebung können die Schüler einer oder mehrerer Klassen eines Jahrgangs einer oder mehrerer Schulen sein. Wenn die Leistungserhebung im Rahmen der Fachschaftsarbeit in allen Parallelklassen eines Jahrgangs erfolgte, ist es nahe liegend, eine Zusammenfassung der Schüler nach Klassen vorzunehmen. Zu den unterschiedlichen Bedingungen in der Grundgesamtheit der Klassen gehört dann das konkrete unterrichtliche Vorgehen des jeweiligen Lehrers. Dies kann Anlass zu fruchtbringenden Diskussionen und Schlussfolgerungen zu Fragen der Unterrichtsgestaltung sein. Allerdings muss beachtet werden, dass die Unterschiede zwischen den Parallelklassen auch noch von anderen Faktoren abhängen, wie dem Klassenklima, der Häufigkeit des Lehrerwechsels und vor allem der leistungsmäßigen Zusammensetzung der Klassen. In der Sekundarstufe I müssen bei der Bildung der Klassen eines Jahrgangs viele Faktoren berücksichtigt werden. Dadurch kann es leicht zu einer ungleichmäßigen Verteilung der leistungsstarken und leistungsschwachen Schüler kommen, die sich im Laufe der Schulzeit weiter vertiefen kann. Einfache Modellrechnungen zeigen (vgl. Eikenbusch; Leuders 2004, S. 60 ff.), dass durch zufällige Einflüsse erhebliche Unterschiede in den Zensurendurchschnitten von Parallelklassen bei einer Vergleichsarbeit auftreten können. Wenn etwa 12 leistungsstarke Schüler zufällig auf drei Parallelklassen aufgeteilt werden, beträgt die Wahrscheinlichkeit nur 6,5 %, das dann in jeder Klasse 4 leistungsstarke Schüler sind.

Wenn es sich um Leistungserhebungen an mehreren Schulen handelt und damit eine Diskussion und Zusammenarbeit der Fachkollegien mehrerer Schulen verbunden ist, sollte nach Möglichkeit eine Zusammenfassung der Fälle nach Schulen erfolgen. Dies setzt allerdings voraus, dass die Bedingungen in den Einzugsgebieten der Schulen und damit die Zusammensetzung der Schülerschaft vergleichbar sind.

Die **Bestimmung der zu testenden Personenmerkmale** sollte durch die Lehrer selbst erfolgen, da sie am besten entscheiden können, in welchen Bereichen ihres beruflichen Könnens aktuell der größte Diskussions- und Veränderungsbedarf besteht. Bei den Merkmalen kann es sich sowohl um spezielle Dispositionen als auch um sehr allgemeine Merkmale wie etwa die Problemlösefähigkeiten der Schüler handeln.

Die Anzahl der betrachteten **Dimensionen** ergibt sich aus dem konkreten Anliegen der Erhebung. Eine Beschränkung auf wenige Dimensionen ist günstig, da dann die Erhebung sehr tiefgründig angelegt und die Auswertung sehr zielgerichtet durchgeführt werden kann.

Eine **Normierung** der Leistungserhebung ist für die Diskussion der Ergebnisse sehr wünschenswert. Bei einer selbst aufgestellten Erhebung ergibt sich eine normorientierte Normierung in der jeweiligen Grundgesamtheit. Dies kann den Schülern zeigen, welchen Rang ihre Leistung in der Grundgesamtheit hat. Eine weitergehende Normierung ist nur möglich, wenn vorliegende externe Testverfahren durch die Lehrer verwendet werden. Sind diese wie etwa durchgeführte Vergleichsarbeiten auf Landesebene normiert, ist ein Vergleich der Leistungen der Schule mit den übrigen Schulen des Landes möglich. Dies setzt voraus, dass die Landesergebnisse schulweise zusammengefasst wurden, wie es etwa bei unseren Vergleichsarbeiten erfolgte.

Günstiger für die Auswertung der Ergebnisse ist eine kriteriumsorientierte Normierung. Nur auf diesem Wege kann erreicht werden, dass Stärken erkannt und Defizite überwunden werden können, die bei einer Normierung an der Grundgesamtheit aller Schüler nicht sichtbar sind. Eine kriteriumsorientierte Normierung kann nur sehr eingeschränkt durch die Lehrer einer Schule erfolgen. Dies müssen Gremien eines Landes auf der Grundlage komplexer Überlegungen und empirischer Ergebnisse leisten.

Da die Auswertung in der Schule erfolgt und eine möglichst große Anzahl von Informationen über die getesteten Personenmerkmale erhalten werden soll, ist es sinnvoll, freie **Antwortformate** zu verwenden. Die zentral bereitgestellten Erhebungsverfahren sollten deshalb auch dieses Format besitzen.

Eine wesentliche **Bedingung bei der Durchführung** dieser Leistungserhebungen ist es, dass sie unvorbereitet für die Schüler erfolgen. Nur so erhält man ein reales Bild von dem tatsächlichen Leistungsstand der Schüler. Die Anforderungen und in einige Fällen auch die Aufgaben sollten den Schüler aber im Prinzip bekannt sein. So können die Schüler bei unserem Konzept des sicheren Wissens und Könnens durchaus alle Aufgaben, die sie stets sicher beherrschen sollen, kennen.

Die Häufigkeit dieser Leistungserhebungen kann nicht zentral reglementiert werden. Es wird sich kaum ein Turnus von Erhebungen zu den gleichen Personenmerkmalen einstellen. Sinnvoll ist eine Konzentration auf einen Bereich von Leistungsdispositionen, so dass alle Bereiche im Laufe des Arbeitslebens eines Lehrers nacheinander abgearbeitet werden und kaum öfter als zweimal vorkommen dürften.

Es gibt keinen oder nur einen losen rechtlichen Rahmen für diese Erhebungen. Die Schüler sollten aber **partizipatorisch** in das Anliegen und Auswertung der Erhebung einbezogen werden, damit sie sich motiviert und engagiert daran beteiligen, obwohl es keine Zensuren gibt. Eine weitere wichtige Bedingung ist, dass diese Leistungserhebungen nicht von außen vorgegeben oder gar erzwungen, sondern nur angeregt werden können. Wenn die Lehrer nicht von sich aus bereit sind, die Ergebnisse ihre Arbeit zu überprüfen, werden die Resultate der Erhebungen kaum den realen Zuständen entsprechen, da es immer unlautere Möglichkeiten gibt, die Ergebnisse in gewünschter Richtung zu verändern. Das eigentliche Ziel, eine offene und nachhaltige Diskussion der Ergebnisse, wird so ebenfalls kaum erreicht.

Wenn die Lehrer die Erhebungsverfahren selbst konzipieren, sind die **Gütekriterien** nur schwer einzuhalten. Dazu wäre ein erheblicher Aufwand bei der Entwicklung des Erhebungsverfahrens nötig, den Lehrer kaum leisten können. Deshalb sollten hinreichend viele Verfahren für möglichst spezielle Bereiche von Leistungsdispositionen vorhanden sein, die den Gütekriterien genügen.

Die **Methoden der Datenauswertung** richten sich danach, ob selbst entworfenen oder fertige Verfahren eingesetzt werden. Eine wesentliche Quelle für Überlegungen zu Ursachen und Konsequenzen ist die Analyse der tatsächlichen Schülerantworten insbesondere der fehlerhaften. Die Leistungserhebungen sollten in der Regel nicht bewertet werden, da das Ziel sowohl eine Überprüfung der Leistungen der Schüler als auch der Lehrer ist. Eine Zensierung kann im Einvernehmen mit den Schülern erfolgen.

Leistungserhebungen mit dieser Funktion sollten ein Hauptfeld der empirischen Untersuchungen in der **didaktischen Forschung** sein. In enger und vertrauensvoller Zusammenarbeit mit den Lehrern können die Wissenschaftler sich an der Erfassung der Bedingungen der Unterrichtsprozesse, der Erstellung und Auswertung von Erhebungsverfahren und den Diskussionen zu den Schlussfolgerungen beteiligen und daraus vielfältige Informationen zu den Gesetzmäßigkeiten der Entwicklung psychischer Dispositionen unter Unterrichtsbedingungen gewinnen. Es können die Aktivitäten der Lehrer sehr gut mit wissenschaftlichen Untersuchungen zu neuen Inhalten und Methoden der Unterrichtsgestaltung verbunden werden.

2.2.4. Funktionen in der Entwicklung von Rahmenbedingungen von Schule und Unterricht

In diesem Bezugssystem können folgenden Teilfunktionen unterschieden werden:

- Funktion der Ermittlung empirischer Vergleichsdaten zur Fundierung von bildungspolitischen Entscheidungen
- Funktion der Ermittlung von Informationen für Schulaufsichtsbehörden und Schulleitungen als Entscheidungshilfe bei Maßnahmen zur Schulentwicklung
- Funktion der Überprüfung von Standards und der Entwicklung von kriteriumsorientierten Diagnoseverfahren

Da es eine Reihe von Unterschieden in den Merkmalen von Leistungserhebungen mit diesen Zielen gibt, sollen sie extra betrachtet werden.

Funktion der Ermittlung empirischer Vergleichsdaten zur Fundierung von bildungspolitischen Entscheidung auf Staats- und Bundesländerebene

Beispiele für solche Erhebungen sind die internationalen Vergleichsstudien TIMSS und PISA, aber auch die nationalen Studien MARKUS und LAU. Es geht dabei um Entscheidungen, die auf Landes- oder Bundesebene getroffen werden müssen und sich in Beschlüssen der KMK, Schulgesetzen der Länder oder Verordnungen niederschlagen. Es kann sich auch um grundsätzliche Richtungsentscheidungen handeln, wie etwa den Ausbau von Ganztagsangeboten oder die Entwicklung von Bildungsstandards.

In diesen Untersuchungen können nur wenige **Bedingungen** des konkreten Unterrichtsverlaufs erfasst werden, wie etwa generelle Einschätzungen zur Computernutzung oder zur Unterrichtsqualität aus Schülerperspektive. Es ist aber für einige Entscheidungen sinnvoll, allgemeine Bedingungen wie das Geschlecht, den Migrationshintergrund oder bestimmte soziale und familiäre Bedingungen zu erfassen.

Die originäre **Grundgesamtheit** sind die Schüler einer bestimmten Altersgruppe eines oder mehrerer Staaten bzw. Bundesländer. Es erfolgt in der Regel eine Gliederung dieser Grundgesamtheit nach Staaten, Staatengruppen oder Bundesländern. Dies bedeutet, dass alle politischen Gebilde als vergleichbar angesehen werden und von sämtlichen Besonderheiten abgesehen wird. Die gemeinsamen Merkmale der Elemente der Grundgesamtheit reduzieren sich dadurch praktisch auf die Tatsachen, dass Schüler hauptsächlich im Klassenverband (in einigen Ländern noch in erheblichem Maße außerhalb der Schule) von mehr oder weniger für das Fach ausgebildeten Lehrern unterrichtet werden und sich dabei mit etwa den gleichen mathematischen Themengebieten beschäftigen.

Bei den jüngsten internationalen Erhebungen TIMSS, PISA und IGLU wurde die Leistung der Schüler in der Domäne Mathematikunterricht mit nur einem **Personenmerkmal** erfasst, das z. B. bei OECD/PISA als „mathematical literacy“ bezeichnet wird. Die Erhebungen sind also in Bezug auf den Mathematikunterricht eindimensional. Bei früheren Untersuchungen wie etwa der IEA-Studie von 1970 wurden neben der Gesamtpunktzahl auch die Teilpunktwerte für 9 bis 11 Teilgebiete in den jeweiligen Populationen ausgewiesen (Postlethwait 1968).

Die **Normierung** erfolgt bei allen internationalen Untersuchungen und den daran orientierten nationalen Erhebungen bisher mit Bezug auf die Gesamtpopulation. Es erfolgt oft eine Transformation der Skalen, so dass der Mittelwert 500 und die Streuung 100 beträgt. Nach Helmke und Hosenfeld (2004, S. 60) besteht aber unter den Bildungsforschern Konsens darüber, dass verteilungsorientierte Aussagen heute nicht mehr ausreichen, sondern kriteriale Aussagen zu inhaltlich definierten Abstufungen von Kompetenzen erwartet werden. Bei den großen internationalen Studien TIMSS und PISA wurde als ein Ansatz in diese Richtung versucht, nach der Datenerhebung eine inhaltliche Deutung der Aufgaben einer Schwierigkeitsstufe, die relativ willkürlich definiert wurden, vorzunehmen.

Um angesichts der jeweils sehr großen Zahl getesteter Schüler überhaupt eine rationelle Auswertung vornehmen zu können, wird bei den meisten Items ein gebundenes **Antwortformat** verwendet. Bei freien Antwortformaten erfolgt bei der Auswertung eine Codierung mit richtig oder falsch. Die damit verbundenen Nachteile des gebundenen Antwortformates müssen zwangsläufig in Kauf genommen werden.

Um die Vergleichbarkeit der Ergebnisse zu gewährleisten, muss man bei allen Untersuchungen darauf achten, dass die Schüler auf die Erhebungen nicht speziell **vorbereitet** werden. Ein großer Teil der Aufgaben muss geheim gehalten werden, um sie bei späteren Erhebungen zu Vergleichszwecken verwenden zu können.

Die **Durchführungsbedingungen** müssen streng reglementiert und kontrolliert werden, um die Durchführungsobjektivität der Erhebung zu gewährleisten.
Es ist wenig sinnvoll, solche Leistungserhebungen in kurzen Abständen zu wiederholen, da die Wirkung von Veränderungen der Rahmenbedingungen auf Landesebene sich erst nach einer gewissen Anzahl von Jahren einstellen dürfte. Damit erklärt sich u. E. auch der Verzicht Deutschlands auf eine zusätzliche Beteiligung neben PISA an der neuen internationalen Studie TIMSS 2007, der vierten kombinierten internationalen Schulleistungsstudie der IEA für die Bereiche Mathematik und Naturwissenschaft[4].

Eine **Partizipation** der Beteiligten erfolgt auf der Ebene der Staaten bzw. Ländern. Die Staaten haben z. B. die Aufgaben und weiteren Analysemethoden gemeinsam entwickelt. Auf dieser Ebene erfolgen dann auch die wesentlichen Auswertungen und Veränderungen. Schulen, Lehrer oder Schüler können bereits aus rein technischen Gründen an der Erhebung nicht partizipatorisch beteiligt werden, da die Auswahl zufällig erfolgt. Sie haben aber auch keine persönlichen Konsequenzen aus denen ihnen erzielten Ergebnissen zu befürchten. Deshalb ist eine Kritik an den Studien unter diesem Aspekt auch nicht angebracht.

Die Absicherung der **Gütekriterien** ist ein wichtiges Anliegen bei der Entwicklung der Erhebungsmethoden. Um in der zur Verfügung stehenden begrenzten Testzeit ein möglichst breites Spektrum von Items einsetzen zu können, werden oft spezielle Methoden verwendet. So können etwa Items auf verschiedene Testhefte verteilt werden, die untereinander durch Ankeritems verbunden sind. Unter der Voraussetzung, dass die Auswahl der Schüler aus der Grundgesamtheit zufällig erfolgt und die Rahmenbedingungen für alle Unterrichtprozesse als im Wesentlichen gleich angesehen werden, kann dann eine Zusammenfassung der Teilergebnisse zu einem Gesamtbild erfolgen.

Die **Methoden der Auswertung der Daten** richten sich nach den verwendeten Testmodellen. Wird als Testmodell das Rasch-Modell verwendet, ergibt sich nur ein Parameter zur Beschreibung der mathematischen Leistungsfähigkeit aller Schüler eines Landes. Da die Adressaten vor allem Bildungspolitiker sind, müssen sich die Analysen und Schlussfolgerungen auf einer sehr allgemeinen Ebene bewegen. So ist etwa ein wesentliches Ergebnis von PISA, dass in Deutschland die Bildungschancen in weit höherem Maße vom sozialökonomischen Status der Eltern abhängen, als in anderen Ländern. Daraufhin wurden Maßnahmen eingeleitet, um die Bildungsgerechtigkeit zu erhöhen.

Eine Bewertung der Schülerleistungen ist bei diesen Erhebungen nicht angebracht.

Für die **didaktische Forschung** sind solche Leistungserhebungen als Bestandteile empirischer Forschung nicht geeignet. Dies zeigen z. B. die kritischen Analysen von Meyerhöfer (2004b) und Bender (2005). Es beginnt schon damit, dass mit den gebundenen Aufgabenformaten wesentliche Komponenten der mathematischen Bildung wie etwa das Problematisieren von Sachsituationen, die

[4] Obwohl es sich um die vierte Studie handelt, wird von der IEA die Bezeichnung TIMSS (Third ….) beibehalten, da sich diese Bezeichnung quasi als ein Markenzeichen der IEA verbreitet hat.

auch in dem Konzept der „mathematical literacy“ enthalten sind, gar nicht erfasst werden können. Das verwendetet Testmodell ist mit didaktischen Belangen kaum vereinbar und schließlich verhindert die notwendige Geheimhaltung von Aufgaben eine öffentliche wissenschaftliche Analyse der Testinstrumente. Didaktiker sollten sich deshalb nur als wissenschaftliche Berater an den Studien beteiligen.

Da die konkreten Unterrichtsbedingungen nicht erfasst werden, bewegen sich die Schlussfolgerungen für den Mathematikunterricht in einem sehr allgemeinen Rahmen.

Als z. B. Anfang 1997 die Ergebnisse von TIMSS-2 veröffentlicht wurden, wurde im Vorstand der GDM, dem ein Autor dieses Buches zur damaligen Zeit angehörte, über die Reaktion auf die Ergebnisse diskutiert. Die damals schon berechtigten Zweifel an den verwendeten Messinstrumenten und der Aussagekraft der Daten wurden zurückgestellt und die Gelegenheit genutzt, um Forderungen zur Veränderung des Mathematikunterrichts aufzustellen, die … schon seit langem erhoben und mit vielen eigenen Vorschlägen auch hinlänglich konkretisiert, in der Breite des Unterrichts bisher aber noch unzureichend realisiert worden sind.“ (Erklärung 1997, S. 37). Im Ergebnis entstand dann z. B. das sehr erfolgreiche SINUS-Projekt.

Als eine Folge solcher Leistungserhebungen können sich eine Reihe von Impulsen für die didaktische Forschung ergeben, wie von Neubrand (2005) herausgestellt wurden. Dabei handelt es sich vor allem um theoretische Probleme wie die Strukturierung mathematischer Leistung oder die Typisierung von Aufgaben.

Funktion der Ermittlung von Informationen für regionale Schulaufsichtsbehörden und Schulleitungen als Entscheidungshilfe bei Maßnahmen zur Schulentwicklung

Es werden hier nur Informationen über die Leistungen im Mathematikunterricht betrachtet.

Das Ziel der Informationsgewinnung sollte kein Selbstzweck sein, sondern im Sinne des Bezugssystems zur Beeinflussung oder Veränderung von Rahmenbedingungen für den Unterricht führen. Dazu gehören die Organisation des schulischen Lebens, die Arbeit des Schulleiters, die Arbeit der Fachkonferenzen, die Gestaltung der außerschulischen Beziehungen u. a.

Nach Peek (2001) sollten zentrale administrierte Vergleichsuntersuchungen weder der Überprüfung der individuellen Leistungen einzelner Schülerinnen und Schüler noch der Kontrolle einzelner Lehrkräfte dienen. Sie sollten weder die Grundlage der individuellen Entscheidungen über Berechtigungen noch ein Element der Personalbeurteilung sein. Ihr Hauptziel ist es, „eine vergleichende Zusammenschau über die Situation der Schulen im eigenen Verantwortungsbereich zu gewinnen und die Ergebnisse für bildungsplanerische Konsequenzen, zum Beispiel für die Entwicklung neuer curricularer Konzeptionen oder die Lehreraus- und -fortbildung zu nutzen. Primäre Zielsetzung externer Leistungsmessungen ist damit die Bereitstellung von Steuerungswissen.“ (S. 330)

Durch entsprechende Maßnahmen an den Schulen können die Prozesse des Bezugssystems der zweiten Funktion, die Entwicklung des beruflichen Könnens von Lehren und Fachkollegien, durch zentrale Leistungserhebungen angeregt und befördert werden. Außerdem ergeben sich weitere Orientierungspunkte für den Stand der eigenen Qualitätsentwicklung, worauf auch Peek (2001) hinweist.

Für die Tätigkeit der Schulaufsichtsbehörden setzt dies allerdings voraus, dass dort genügend personelle Kapazitäten für Aktivitäten an Schulen vorhanden sind. Bei den Vergleichsarbeiten in Mecklenburg-Vorpommern in den Jahren 1998 und 1999 wurden zwar auch die Ergebnisse aller Schulen im Bildungsministerium erfasst, aber es erfolgte aus Kapazitätsgründen keine Kontaktaufnahme mit ausgewählten Schulen.

Eine Erfassung der **Unterrichtsbedingungen** kann sich auf den Bereich beschränken, der durch Schulbehörden und Schulleitungen beeinflusst werden kann, wie Art und Inhalt von schuleigenen Lehrplänen, Programmen zur Fachschaftsarbeit oder zur Fortbildung, die Nutzung der Unterrichtszeit u. a.

Die **Grundgesamtheit** sollte die Menge aller Schulen eines Aufsichtsbereiches (z. B. Schulamtes) und nicht der Klassen der Schulen sein. Eine differenzierte Bewertung der Unterschiedlichkeit von Klassen kann nur in Verantwortung des Schulleiters auf Schulebene erfolgen, da zu viele Faktoren diese Unterschiede bedingen.

Die internationalen Studien können für die Ziele dieser Leistungserhebungen nicht verwendet werden, da keine Vollerhebung an den Schulen eines Aufsichtsbereiches bzw. eine repräsentative Stichprobenauswahl stattfand.

Bei der **Bestimmung der zu testenden Personenmerkmale** sind zwei gegensätzliche Aspekte zu beachten. Von den Schulbehörden wird in der Regel erwartet, dass mit der Leistungserhebung eine möglichst allgemeine Aussage über den Mathematikunterricht möglich ist. Dies führt tendenziell zu einer Beschränkung auf sehr wenige oder möglichst nur ein einziges Personenmerkmal, das für die gesamte mathematische Leistungsfähigkeit steht, wie es etwa bei den internationalen Studien der Fall ist. Andererseits sollten die Ergebnisse der Erhebung aber auch für spezielle Aktivitäten zu einzelnen Bereichen des Unterrichts verwendbar sein, um etwa Diskussionen zu konkreten Veränderungen im Unterricht anzuregen oder auch Handreichungen oder Fortbildungen auf Landesebene zu unterstützen. Dies führt tendenziell dazu, möglichst spezielle Personenmerkmale auszuwählen, mehrere zu berücksichtigen und diese tiefgründig zu untersuchen. Es ist für die Leitungstätigkeit eines Schulleiters außerdem viel günstiger, wenn er Aussagen zu konkreten Leistungsbereichen erhält. Eine pauschale Bewertung mit einer Durchschnittszahl oder ein eindimensionales Ranking hilft wenig. Wenn gleichzeitig mehrere Leistungsbereiche getestet werden, so zeigen sich in der Regel sowohl bestimmte Stärken als auch Schwächen, wie unsere Erfahrungen mit den Vergleichsarbeiten bestätigen. In Abwägung beider Tendenzen plädieren wir für eine größere Anzahl von **Dimensionen** und eine tiefgründige Analyse sowie einen bewussten Verzicht auf eine pauschale Gesamtbewertung.

Eine **Normierung** kann bezogen auf die Schülerpopulation der betrachteten Schulen erfolgen, da es nur um die Verteilung der Leistungen geht und man sich in der Regel nur auf Extremgruppen von Schulen mit besonders hohen und mit besonders geringen Schülerleistungen konzentrieren kann. Ein Vergleich mit einer größeren Population wie die Schüler aller Bundesländer oder gar aller PISA-Länder bringt für die konkrete Arbeit wenig Nutzen. Liegt man z. B. bei diesem Vergleich mit der Mehrzahl der Schulen über dem Durchschnitt der größeren Population, kann man nicht die Folgerung ableiten, kaum mehr etwas tun zu müssen. Liegt man weit unter dem Durchschnitt, ist es nicht angebracht zu resignieren und nichts oder besonders viel zu tun. Solche Vergleiche können bestenfalls eine innere Zufriedenheit und Unzufriedenheit der Mitarbeiter der Schulbehörden erzeugen. In der täglichen Arbeit geht es ihnen wie einem Lehrer, die differenzierte Arbeit ist ein grundsätzliches Erfordernis, unabhängig vom absoluten Bezug. Die gleichen Argumente ergeben sich hinsichtlich einer kriteriumsorientierten Normierung, die für diese Zwecke ebenfalls nicht erforderlich ist.

Als **Aufgabenformat** sollte vor allem das den Schülern aus dem Unterricht vertraute freie Antwortformat gewählt werden, wobei auch durchaus Aufgaben mit gebundenem Format vorhanden sein können. Dies führt allerdings zum Problem eines erhöhten Aufwandes bei der Auswertung dieser Arbeiten wozu es jeweils spezielle Auswertungsgruppen im Lande geben müsste. Wie aber die Erfahrungen mit den Vergleichsarbeiten in unserem Land zeigen, sind diese Probleme durchaus mit einem relativ geringen finanziellen Aufwand unter Einbeziehung von Hilfskräften und der Kapazitäten der Universitäten und Landesinstitute lösbar.

Um die Vergleichbarkeit der Schulen zu gewährleisten, ist unbedingt zu sichern, dass die Erhebungen **unvorbereitet** für die Schüler erfolgen. Dem muss nicht widersprechen, dass die Anforderungen in diesen Arbeiten Schülern und Lehrern durchaus bekannt sind. Es würde diesem Anliegen entsprechen, wenn bei der Auswahl der Personenmerkmale eine bewusste Eingrenzung auf bestimmte Bereiche erfolgt, die vorher nicht bekannt gegeben werden.

Die **Durchführungsbedingungen** dieser Leistungserhebungen sollten wie bisher auch üblich den Bedingungen von zentralen Abschlussprüfungen entsprechen.

Bei der Vorbereitung der Leistungserhebungen und der damit möglicherweise verbundenen Befragungen an den Schulen sollten die Schulen als Elemente der Grundgesamtheit **partizipatorisch** einbezogen werden. Sie sollten Art und Umfang der Erhebung mitdiskutieren können.

Da sich eine Beeinflussung von Schulen und die damit verbundenen Veränderungen erst langfristig auf die Leistungen im Mathematikunterricht auswirken werden, sollten diese Leistungserhebungen nicht in einem regelmäßigen **Turnus**, sondern nur in größeren Abständen erfolgen. Es ist nicht sinnvoll, daraus eine jährlich stattfindende Aktion zu machen wie es leider gegenwärtig in vielen Bundesländern der Fall ist.

Es sollte versucht werden die **Gütekriterien** von Testverfahren möglichst gut zu erfüllen. Dazu würde es gehören, eine Pilotierung der Aufgaben vorzunehmen und entsprechende Untersuchungen zur Validität und Reliabilität anzustellen.

Bei der **Auswertung der Daten** sollte man sich auf die den meisten Lehrern bekannten Methoden der klassischen Testtheorie beschränken. In der Regel sollte keine Bewertung der Schülerleistungen erfolgen, da die Schüler nicht für Mängel in den Rahmenbedingungen der Schule verantwortlich gemacht werden können.

Für die **didaktische Forschung** ergeben sich mögliche Projekte, in dem man sich an der Vorbereitung und Auswertung der Erhebungen beteiligt. So wäre die Auswertung der erhobenen Schülerantworten insbesondere bei offenen Aufgabenformate mit den Methoden der Fehlergruppenanalyse möglich (vgl. Kapitel 3). Weiterhin könnten Untersuchungen zu ausgewählten Unterrichtsbedingungen für diese Ergebnisse erfolgen. Dazu ist es aber erforderlich, dass man sich auf Stichprobenschulen beschränkt und dort weitere Untersuchungen durchführt.

Die eigentliche Planung und Durchführung dieser Leistungserhebungen sowie die inhaltliche Verantwortung sollte aber bei den Landesbehörden bleiben und nicht als ein Forschungsprojekt an Wissenschaftler übergeben werden.

Funktion der Überprüfung von Standards und der Entwicklung von kriteriumsorientierten Diagnoseverfahren für die Hand des Lehrers

Ziel dieser Leistungserhebungen ist es nicht, Einschätzungen von Schulen vorzunehmen oder bildungspolitische Konsequenzen abzuleiten. Es geht um die Konkretisierung von Bildungsstandards und die Entwicklung von Testverfahren für die Hand des Lehrers. Wir stehen bei beiden Zielbereichen in Deutschland erst am Anfang der Entwicklung. Die gegenwärtige Form der Bildungsstandards für Mathematik ist noch weit davon entfernt, konkrete Vorgaben im Sinne von Standards für das jeweilige Abschlussniveau zu setzen. Auch die darin enthaltenen Kompetenzmodelle bedürfen einer wesentlichen Weiterentwicklung (Sill 2006).

Helmke und Hosenfeld (2004) haben in sehr deutlicher Weise die gegenwärtigen Desiderata herausgestellt und einen Kreisprozess zur stufenweisen Entwicklung von Standards und Testaufgaben vorgeschlagen. Es sollten in spiralförmiger Weise ausgehend von den aktuellen Kompetenzmodellen und Bildungsstandards Testaufgaben erzeugt werden, deren empirische Bewährung zu einer Weiterentwicklung der Bildungsstandards und auch der Kompetenzmodelle führen. (S. 62)

Leistungserhebungen mit diesen Zielen sind eine notwendige Bedingung für die Leistungserhebungen im Sinne der zweiten Funktion, da Lehrer für ihre schulinternen Evaluationen einen kriterialen Vergleichsmaßstab und entsprechende Messinstrumente benötigen.

Helmke und Hosenfeld (2004) haben weiterhin betont, dass die Standardsetzung keine Aufgabe der empirischen Bildungsforschung ist, sondern eine fachdidaktisch basierte bildungspolitische Entscheidung darstellt. In die aktuelle Forschungsstruktur sind Didaktiker völlig unzureichend eingebunden, so dass eine solche Entwicklung kaum zu gewährleisten ist. Im Juni 2004 wurde an der Humboldt-Universität Berlin von der KMK ein Institut für Qualitätsentwicklung gegründet, das von allen Bundesländern mitfinanziert wird. Dieses Institut hat das Ziel, nationale Bildungsstandards weiterzuentwickeln, sie zu normieren, ihre Erreichung zu überprüfen und ihre Implementation wissenschaftlich zu begleiten. Dazu gehören auch das Formulieren von Kompetenzmodellen und die Erarbeitung von computergestützten Test-, Auswertungs- und Rückmeldesystemen. Der Leiter des Instituts und anscheinend auch die Mehrzahl der Mitarbeiter stammen allerdings aus dem Bereich der Bildungsforschung und als Forschungsgebiete werden lediglich die pädagogisch-psychologische Diagnostik und die „Lehr-/Lernforschung im Schnittbereich zwischen Fachdidaktiken, Erziehungswissenschaft und Psychologie“ angesehen.

Die zu entwickelnden Testverfahren für die Hand des Lehrers sollten Leistungserhebungen im Sinne der zweiten Funktion ermöglichen. Die bisher entwickelten Testverfahren entsprechen bisher nicht den dazu notwendigen Anforderungen. Die wenigen vorhandenen standardisierten psychologischen Tests sind nicht geeignet, da sie nur mit einem festgelegten Itemsatz einsetzbar und nicht kostenlos an allen Schulen verfügbar sind. Sie sind weiterhin wie auch die meisten der gegenwärtig verwendeten Testverfahren nur verteilungsorientiert normiert. Es muss ein neuer Typ von frei zugänglichen Testverfahren entwickelt werden, der u. a. eine zufällige Auswahl von gleichwertigen Items ermöglicht.

Die Konkretisierung von Bildungsstandards und die Entwicklung von kriteriumsorientiert normierten Testverfahren können durch die gleichen Leistungserhebungen erfolgen, die die gleichen im Folgenden angegebenen Merkmale besitzen.

Die Erfassung der unterrichtlichen **Bedingungen** ist aus zwei Gründen erforderlich. Wenn den Lehrern ein bestimmter Standard vorgegeben und damit verbunden ein kriteriumsorientiertes Testverfahren zur Verfügung gestellt wird, sollte ihnen auch mitgeteilt werden, unter welchen Bedingungen sich diese Normen erreichen lassen. Dazu gehören unter anderem der notwendige Zeitaufwand für Erarbeitungs- und Übungsstunden sowie auch ein geeignetes didaktisch-methodisches Vorgehen. Für die schrittweise Festlegung von Standards ist es weiterhin erforderlich zu wissen, welcher Aufwand hinter der Realisierung bisher festgelegter Standards steckt, da in der Summe ein bestimmter Arbeits- und Zeitaufwand von Lehrern und Schülern nicht überschritten werden kann.

Als **Grundgesamtheit** sollte die Menge aller Schüler der Bundesrepublik gewählt werden, da es sich bei den Bildungsstandards um nationale Standards handelt.

Als zu testenden **Personenmerkmalen** ist es unumgänglich, möglichst kleine und abgegrenzte Bereiche psychischer Dispositionen zu wählen, damit für den unterrichtenden Lehrer eine zielgerichtete Überprüfung des Vorgehens in einzelnen Teilgebieten möglich ist und dazu ein möglichst großes Repertoire von Diagnoseverfahren vorhanden ist.

Wenn reale Aufgabenstellung aus dem normalen Unterricht verwendet werden, ist es unumgänglich das jeweils eine größere Anzahl von **Dimensionen** betrachtet wird.

Das Ziel der Entwicklung der Testverfahren ist eine kriteriumsorientierte **Normierung**. Als Zwischenschritt sollten aber auch verteilungsorientierte Normierungen zugelassen werden.

Das **Antwortformat** sollte bei diesen Aufgaben in der Regel nicht gebunden sein, damit aus den Schülerantworten möglichst viele diagnostische Informationen gewonnen werden können.

Je nach Einsatz der Diagnoseverfahren, sind verschiedene Formen der **Vorbereitung** denkbar. Tests für Basiskompetenzen sollten generell unvorbereitet erfolgen, da dies den Intentionen dieser Kompetenzen entspricht. Es sollte aber auch möglich sein, die Diagnoseverfahren für normale Klassenarbeiten am Abschluss eines Stoffgebiets zu verwenden und damit dann auch eine Vorbereitung der Schüler zu ermöglichen.

Die Leistungserhebungen müssen die **Gütekriterien von Testverfahren** möglichst gut zu erfüllen. Eine Einbeziehung der Lehrer in die Entwicklung der Testverfahren und damit auch die Durchführung der Leistungserhebung ist unbedingt erforderlich, da schon bei der Konstruktion der Aufgaben das zu erreichende Niveau nur durch Mitarbeit erfahrener Lehrer grob bestimmbar ist.

Die Auswertung der Daten der Leistungserhebungen muss nicht transparent für Lehrer und Schüler erfolgen, da es sich um Erhebungen handelt, die in den Entwicklungsprozess von Testverfahren eingeordnet sind. Auf eine Bewertung der Schülerleistungen verbietet sich aufgrund des experimentellen Charakters der Leistungserhebung.

Die Leistungserhebungen und die damit verbundene Bestimmung von Bildungsstandards und Entwicklung von Testverfahren erfordert notwendigerweise eine wissenschaftliche Kooperation von Fachdidaktiker, Bildungsforschern und Psychologen.

2.2.5. Vergleiche von Merkmalen bei Leistungserhebungen mit unterschiedlichen Funktionen

Angesichts des in der Regel sehr hohen Aufwandes, der insbesondere bei zentralen Erhebungen mit der Vorbereitung, Durchführung und Auswertung der Erhebung verbunden ist, wäre es wünschenswert, dass eine Erhebung gleichzeitig mehrere Funktionen erfüllen kann. Die in den vorherigen Abschnitten vorgenommenen Analysen der Merkmale von Leistungserhebungen mit den jeweiligen Funktionen zeigten jedoch, dass eine solche Multifunktionalität von Leistungserhebungen in der Regel nicht möglich ist, da sich die Merkmale der Erhebungen mit den jeweiligen Funktionen in der Summe erheblich unterscheiden. Einige wesentliche Unterschiede sollen noch einmal zusammenfassend genannt werden.

Während für die Ermittlung empirischer Vergleichsdaten auf nationaler Ebene bzw. für regionale Schulbehörden nur solche *Bedingungen* erfasst werden müssen, die auch durch das Land bzw. die regionalen Behörden beeinflusst werden können, müssen bei den Leistungserhebungen mit den übrigen Funktionen die Bedingungen des Unterrichtverlauf möglichst genau bekannt sein.

Die *Grundgesamtheit* sollten einmal nur Schüler einer oder mehrerer Klassen (bei Leistungserhebung zu Entwicklung des beruflichen Könnens von Lehrern) oder auch die Schüler der ganzen Bundesrepublik (bei Leistungserhebung zu Entwicklung von kriteriumsorientierten Verfahren) sein.

Die *Anzahlen der zu erfassenden Personenmerkmale* reichen von einem einzigen Merkmal bei Leistungserhebungen auf Länderebene bis zu einer größeren Anzahl von Merkmalen bei Leistungserhebungen zur Leistungsbewertung von Schüler.

Leistungserhebungen zur Leistungsbewertung von Schülern, zur Ermittlung empirischer Vergleichsdaten auf Länderebene und regionaler Ebene können *normorientiert* normiert werden, während man für die übrigen Leistungserhebungen *kriteriumsorientiert* normierte Tests benötigt.

Die *Antwortformate* bei Leistungserhebungen zur Leistungsbewertung, zur individuellen Diagnostik und zur Entwicklung des beruflichen Könnens von Lehrern sollten offen sein, während sie für die übrigen Arten von Leistungserhebungen durchaus gebunden sein können.

Alle Formen von Leistungserhebungen bis auf die Erhebungen zur Leistungsbewertung von Schülern sollten *unvorbereitet* erfolgen. Auch eine *Bewertung* der Schülerleistungen mit einer Note ist nur für Leistungserhebungen sinnvoll, die die Funktion der Leistungsbewertung haben.

Literaturverzeichnis

Bos, W. u. a. (2003): Erste Ergebnisse aus IGLU - Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. – Münster. Waxmann, 2003

Eikenbusch, G.; Leuders, T. (2004): Lehrer-Kursbuch Statistik - Alles über Daten und Zahlen im Schulalltag. – Berlin. Cornelsen, 2004

Heller, K. A. ; Hany, E. A. (2001): Standardisierte Schulleistungsmessungen. – In: Weinert, F. E. (Hrsg.): Leistungsmessungen in Schulen. – Weinheim: Beltz Verlag, 2001, S. 87-101

Helmke, A.; Hosenfeld, I. (2003): Vergleichsarbeiten (VERA). Eine Standortbestimmung zur Sicherung schulischer Kompetenzen. – In: SchulVerwaltung Hessen, Rheinland-Pfalz, Saarland 2003, H. 1, S. 10-13 und 2003, H. 2, S. 41-43

Helmke, A.; Hosenfeld, I. (2004): Vergleichsarbeiten - Standards - Kompetenzstufen: Begriffliche Klärung und Perspektiven. – In: Jäger, R. S.; Frey, A.; Wosnitza, M. (Hrsg.): Lernprozesse, Lernumgebungen und Lerndiagnostik. Wissenschaftliche Beiträge zum Lernen im 21. Jahrhundert. - Landau: Verlag Empirische Pädagogik, S. 56-75

Helmke, A.; Jäger, R. S. (Hrsg.) (2002): Das Projekt MARKUS – Mathematik-Gesamterhebung Rheinland-Pfalz - Kompetenzen, Unterrichtsmerkmale, Schulkontext. – Landau: Verlag Empirische Pädagogik, 2002

Herwig, Ch.; Böttner, J. (2001): Vergleichsuntersuchungen zum Mathematikunterricht auf der Basis zentraler Prüfungen. – In: Kaiser, G. u. a. (Hrsg.): Leistungsvergleiche im Mathematikunterricht - Ein Überblick über aktuelle nationale Studien. – Hildesheim. Franzbecker, 2001, S. 117-138

Ingenkamp, K. (1997): Lehrbuch der Pädagogischen Diagnostik - Studienausgabe. – 4. Aufl. – Weinheim. Beltz Verlag, 1997

Kaiser-Meßmer, G.; Blum, W. (1993): Einige Ergebnisse von vergleichenden Untersuchungen in England und Deutschland zum Lehren und Lernen von Mathematik in Realitätsbezügen. – In: JMD 14 (1993) 3/4, S. 269-305

Lind, G. (2003): Jenseits von PISA – Für eine neue Evaluationskultur. – URL: http://www.uni-konstanz.de/ag-moral/hodi/et-evaluation.htm auch in: Pädagogische Hochschule Schwäbisch Gmünd, Hrsg.,

Max-Planck-Institut für Bildungsforschung (2002): Stellungnahme zur Meldung der dpa über die PISA-Ergebnisse der Laborschule Bielefeld und der Helene-Lange-Schule in Wiesbaden vom 13. November 2002

Meyerhöfer, W. (2004 a): Zum Problem des Ratens bei PISA. – In: JMD 25 (2004) 1, S. 62-69

Neuman, R. (2000): Sind gemeine Brüche und Dezimalbrüche zwei verschiedenen Arten von Zahlen oder zwei verschiedenen Schreibweisen für ein und dieselben Zahlen. - In: Der Mathematikunterricht 46 (2000) 2, S. 38-49

Padberg, F.; Bienert, T. (2000): Zur Entwicklung des Bruchzahlverständnisses und der Rechenoperationen mit gemeinen Brüchen innerhalb eines Schuljahres. - In: Der Mathematikunterricht 46 (2000) 2, S. 24-37

Peek, R. (2001): Die Bedeutung vergleichender Schulleistungsmessungen für die Qualitätskontrolle und Qualitätsentwicklung von Schulen und Schulsystemen. – In: Weinert, F. E. (Hrsg.): Leistungsmessungen in Schulen. – Weinheim: Beltz Verlag, 2001, S. 323-335

Postlethwait, T. N. (1968): IEA Leistungsmessung in der Schule - Eine internationale Untersuchung am Beispiel des Mathematikunterrichts. – Frankfurt/Main. Diesterweg, 1968

Rost, J. (2004): Lehrbuch. Testtheorie – Testkonstruktion. – 2. Aufl. – Bern. Huber, 2004

Sill, H.-D. (1997): Funktionen und Zielstruktur des Mathematikunterrichts. - In: Beiträge zum Mathematikunterricht 1997. – Hildesheim: Franzbecker, 1997. S. 466 – 469

Sill, H.-D. (2002): Zur Taxonomie der Ziele des Mathematikunterrichts. – In: Beiträge zum Mathematikunterricht 2002. – Hildesheim: Franzbecker, 2002. S. 459 – 462

Sill, H.-D. (2006): PISA und die Bildungsstandards. – In: Pisa & Co - Kritik eines Programms /Jahnke, Th.; Meyerhöfer, W. (Hrsg.). – Hildesheim. Franzbecker, 2006, S. 293-330

Stern, E.; Hardy, I. (2001): Schulleistungen im Bereich der mathematischen Bildung. In: Weinert, F. E. (Hrsg.), Leistungsmessungen in Schulen. S. 153-168.

Weinert, F. E. (2001a): Vergleichende Leistungsmessungen in Schulen – eine umstrittene Selbstverständlichkeit. – In: Weinert, F. E. (Hrsg.): Leistungsmessungen in Schulen. – Weinheim: Beltz Verlag, 2001, S. 17-31

Woschek, R. (2004): Ein Beitrag zur Diskussion des Rateproblems bei MC-Aufgaben. – In: JMD 25 (2004) 2, S. 149-152

Wottawa, H.; Thierau, H. (1998): Lehrbuch Evaluation. – 2. Aufl. – Bern. Huber, 1998

Zech, F.; Wellenreuther, M. (1992): Konstruktive Entwicklungsforschung: eine zentrale Aufgabe der Mathematikdidaktik. – In: JMD 13 (1992) 2/3, S. 143-198