Manuel Ade-Thurow
Wilfried Bos
Andreas Helmke
Tuyet Helmke
Nina Hovenga
Morena Lebens
Gerlinde Lenske
Detlev Leutner
Giang Hong Pham
Anna-Katharina Praetorius
Friedrich-Wilhelm Schrader
Christian Spoden
Joachim Wirth
(Hrsg.)

Aus- und Fortbildung der Lehrkräfte

in Hinblick auf Verbesserung der Diagnosefähigkeit, Umgang mit Heterogenität und individuelle Förderung

WAXMANN

Manuel Ade-Thurow, Wilfried Bos, Andreas Helmke,
Tuyet Helmke, Nina Hovenga, Morena Lebens,
Gerlinde Lenske, Detlev Leutner, Giang Hong Pham,
Anna-Katharina Praetorius, Friedrich-Wilhelm Schrader,
Christian Spoden, Joachim Wirth (Hrsg.)

Aus- und Fortbildung der Lehrkräfte

in Hinblick auf Verbesserung der Diagnosefähigkeit, Umgang mit Heterogenität und individuelle Förderung

Waxmann 2014
Münster • New York

Die Produktion dieses Materials zum Einsatz in der Lehrerbildung wurde ermöglicht durch

Deutsche Telekom Stiftung

Bibliografische Informationen der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

ISBN 978-3-8309-2773-0

www.waxmann.com
info@waxmann.com

Umschlaggestaltung: Anne Breitenbach, Tübingen
Druck: SDK Systemdruck, Köln
Gedruckt auf alterungsbeständigem Papier, säurefrei gemäß ISO 9706

Printed in Germany

Hintergrund und Ziele des Projekts UDiKom

Zu den im Nachgang zu PISA und PISA E von der KMK identifizierten Handlungsfeldern gehören „Maßnahmen zur Verbesserung der Professionalität der Lehrkräfte, insbesondere im Hinblick auf diagnostische und methodische Kompetenz als Bestandteil systematischer Schulentwicklung".

Das Land Nordrhein-Westfalen, vertreten durch das Ministerium für Schule und Weiterbildung (MSW), war zuständig für die Koordination des Projektes. Das MSW hat Prof. Wilfried Bos mit der Koordination der vorliegenden Publikation beauftragt.

Der Umsetzung dienen die Projekte ProLesen, das eine Stärkung der Lesekompetenz der Schülerinnen und Schüler zum Ziel hat, das Projekt for.mat zur Bereitstellung von Fortbildungskonzeptionen und -materialien zur kompetenz- bzw. standardbasierten Unterrichtsentwicklung sowie die Erarbeitung von Standards und niveaubestimmenden Aufgaben durch das Institut zur Qualitätsentwicklung im Bildungswesen (IQB), und das Projekt UDiKom. UDiKom hat das Ziel, ausgehend von wissenschaftlichen Forschungsergebnissen Lehrkräfte in allen drei Phasen der Aus- und Fortbildung im Hinblick auf ihre diagnostischen Kompetenzen zu stärken, um den Umgang mit Heterogenität zu verbessern und individuelle Förderung zu ermöglichen. Studienbriefe, Materialien und Instrumente sind entwickelt worden für die Bereiche:

- **Individualdiagnostik (Projektleiter: Prof. Joachim Wirth, Ruhr-Universität Bochum)**
  Instrumente und Planung ihres Einsatzes für die eigene Unterrichtsentwicklung als Grundlage für individuelle Förderung
- **Vergleichsarbeiten/Lernstandserhebungen (Projektleiter: Prof. Detlev Leutner, Universität Duisburg-Essen)**
  Verfahren und Interpretation der Ergebnisse als Grundlage für die eigene Schul- und Unterrichtsentwicklung
- **Internationale Schulleistungsstudien (Projektleiter: Prof. Wilfried Bos, Technische Universität Dortmund)**
  Verfahren und Einschätzung der Ergebnisse für die Schul- und Unterrichtsentwicklung
- **Unterrichtsdiagnostik (Projektleiter: Prof. Andreas Helmke, Universität Koblenz-Landau)**
  Instrumente und Verfahren der Unterrichtsdiagnostik, ihr Einsatz und ihre Interpretation der Ergebnisse als Grundlage für die eigene Unterrichtsentwicklung

Die drei Studienbriefe des Moduls 1 sind gleich strukturiert: Sie beschreiben zu den jeweiligen Themen die Zielsetzungen, Bewertungskriterien, testtheoretischen Grundlagen sowie den Anwendungsbereich und zeigen praktische Implikationen auf. Sie haben eine gemeinsame Einleitung, bieten ein gemeinsames Glossar zur Erläuterung von wissenschaftlichen Fachbegriffen und sind über Referenzen und Querverweise miteinander verbunden.

Darüber hinaus wurde an der Ruhr-Universität Bochum eine Datenbank zur Klassifikation von Testinstrumenten erstellt. Weiterhin wurden an der Technischen Universität Dortmund interaktive Trainingsmodule und eine Online-Diskussionsplattform für Blended Learning entwickelt. Als zusätzliche Ergänzung der Studienbriefe des Moduls 1 wurden Foliendatensätze und ein Einführungsfilm erstellt.

Das Modul 2 beschäftigt sich mit der Diagnostik des Unterrichts. Der von Prof. Dr. Andreas Helmke (Universität Koblenz-Landau) und seinem Team entwickelte Studienbrief ergänzt das Standardlehrbuch „Unterrichtsqualität und Lehrerprofessionalität" um ein handlungsorientiertes Werkzeug, das eine kriterienorientierte und evidenzbasierte Reflexion des Unterrichts ermöglichen soll. Konstitutiver Bestandteil dieses Studienbriefes ist ein netzbasiertes Programm zur Selbstevaluation auf Basis selbst erhobener Daten zum Unterricht (Schülerfeedback oder Unterrichtsbeobachtung).

Als Angebot der Wissenschaft an die Praxis erhebt das vorgestellte Material den Anspruch, flexibel als Grundlage in allen Formen der Lehrerbildung in unterschiedlichen Lernkontexten einsetzbar und mit anderen Instrumenten kombinierbar zu sein.

Handlungsleitend bei der Erarbeitung der Studienbriefe, der Instrumente und der ergänzenden Materialien war einerseits der Anspruch nach wissenschaftlicher Korrektheit und andererseits der Anspruch, den Anforderungen unterschiedlicher Zielgruppen und unterschiedlicher Veranstaltungsformate in der Ausbildung und Fortbildung von Lehrkräften gerecht zu werden.

Die Ergebnisse des Moduls 1 stehen allen interessierten Nutzerinnen und Nutzern kostenlos und ohne Zugangsbeschränkungen unter der Adresse www.udikom.de zur Verfügung. Für die Einstellung von Instrumenten in die Datenbank ist ein Anmeldeverfahren vorgesehen.

Unter der Adresse www.unterrichtsdiagnostik.info stehen die Ergebnisse des Moduls 2 allen interessierten Nutzerinnen und Nutzern kostenlos und ohne Zugangsbeschränkungen zur Verfügung.

Inhalt

Individualdiagnostik

Joachim Wirth
Morena Lebens

UDiKom

Aus- und Fortbildung der Lehrkräfte in Hinblick auf Verbesserung der Diagnosefähigkeit, Umgang mit Heterogenität, individuelle Förderung

Individualdiagnostik

Alle im Projekt erstellten Materialien
finden Sie unter
www.udikom.de

1.1 Gegenstand und Zielsetzung

In diesem Kapitel behandelte Fragen:

- *Welchen pädagogischen Ertrag bieten individualdiagnostische Verfahren?*
- *Wodurch zeichnet sich die pädagogische Individualdiagnostik aus?*

Hintergrund

Eine der wichtigsten Aufgaben von Lehrern ist sicherlich die genaue Einschätzung ihrer Schüler, sei es im Bezug auf ihre Fachleistungen, ihre Lernmotivation, ihre sozialen Fähigkeiten oder andere Merkmale, die Lernen beeinflussen können. Diagnostik ist daher ein bedeutsamer Teil der alltäglichen Arbeit im Klassenzimmer. Durch die richtige Einschätzung der Leistungen und Fähigkeiten der Schüler können diese optimal gefördert werden, was wiederum einen enormen Einfluss auf die Motivation der Schüler hat. Falsche Entscheidungen können dagegen die Bildungskarriere eines Schülers sehr erschweren. Darum ist es sehr wichtig, dass pädagogische Diagnostik niemals aus dem Bauch heraus betrieben wird, sondern immer systematisch, geplant und objektiv ist.

Gute pädagogische Diagnostik ist jedoch nicht nur entscheidend für den Schüler. Auch Lehrer profitieren, wenn ihre pädagogischen Entscheidungen auf solider Diagnostik basieren: Jeder im aktiven Schuldienst kennt bspw. die Situation, wenn ein Schüler auf Grund schlechter Zeugnisnoten nicht versetzt wird. Beim Gespräch mit den Eltern sind sich diese dann sicher: Die 5 in Deutsch ist völlig unberechtigt, nicht der Schüler ist zu schlecht, sondern die Klassenarbeiten waren viel zu schwer und überhaupt; im Unterricht wurden doch ganz andere Dinge behandelt als die, die hinterher abgefragt wurden; Schuld ist der Lehrer, der den Schüler doch schon seit dem letzten Schuljahr immer strenger bewertet als den Rest der Klasse. Eine schwierige Situation für jeden Lehrer. Umso wichtiger sind hier gute Argumente dafür, dass der Lehrer den Schüler richtig bewertet hat. Genau diese Argumente liefert eine gute pädagogische Individualdiagnostik.

Ziel der Individual-diagnostik

Ziel der Individualdiagnostik ist es, die Ausprägung verschiedener psychologischer Merkmale genau zu erfassen. Für Lehrer bedeutet dies: durch die Individualdiagnostik lässt sich bspw. einschätzen, wie gut oder schlecht die Leistungen und Fähigkeiten einzelner Schüler sind. Aber es geht nicht nur um Leistung. Gute Individualdiagnostik kann bspw. auch Auskunft darüber geben, wie Schüler selbst ihre Fähigkeiten einschätzen, wie motiviert sie sind, ob sie ängstlich sind oder vieles mehr.

Die Individualdiagnostik kann somit Informationen liefern, die Lehrkräfte benötigen, um bildungsbezogene Entscheidungen für die jeweilige Schülerin/den jeweiligen Schüler begründet treffen zu können. Soll der Schüler versetzt werden? Welche Form der Unterstützung braucht der Jugendliche? Bleibt der Schüler aufgrund falscher Selbsteinschätzung hinter seinen Möglichkeiten zurück? Solche Fragen lassen sich durch eine gute Individualdiagnostik objektiv beantworten, und die daraus gezogenen Konsequenzen lassen sich nachvollziehbar begründen.

In Unterrichtssituationen treffen Lehrkräfte pädagogische Entscheidungen häufig auf Grundlage zufälliger sowie auch geplanter Beobachtungen der Schülerinnen und Schüler (Kliemann, 2008). Für eine Vielzahl pädagogischer Entscheidungen sind solche Beobachtungen auch ausreichend, zumal sie sehr ökonomisch, bspw. während des Unterrichts ohne nennenswerten zusätzlichen Aufwand, durchgeführt werden können. Allerdings müssen sich Lehrkräfte auch darüber bewusst sein, dass derartige Beobachtungen die tatsächlichen Fähigkeiten von Schülerinnen und Schülern häufig nur ungenau abbilden. Daher geben Schrader und Helmke (2001) zu bedenken: „Eine zutreffende Einschätzung des Leistungsstandes ist allerdings eine außerordentlich schwierige Aufgabe, die ohne den Einsatz von professionell entwickelten, am Lehrplan orientierten diagnostischen Instrumenten kaum möglich ist“ (S. 50).

Funktionen der Individual-diagnostik

Individualdiagnostische Verfahren sind daher eine notwendige Ergänzung zu Lehrerbeobachtungen. Dabei kann die Individualdiagnostik verschiedene Funktionen erfüllen. Bspw. können individualdiagnostische Verfahren als Lernausgangsdiagnose Auskunft über den Vorwissensstand oder die Motivation der Schülerinnen und Schüler bieten. Prozessbegleitend kann die Individualdiagnostik innerhalb einer Unterrichtssequenz zur Nachvollziehbarkeit individueller Lernprozesse beitragen. Am Ende der Unterrichtssequenz dient die Individualdiagnostik der Erfassung und Analyse des Lernergebnisses.

Im vorliegenden Studienbriefteil 1 liegt das Hauptaugenmerk auf der Diagnose von Schulleistung und den Merkmalen, die Schulleistung beeinflussen können (schulleistungsrelevante Merkmale). Dabei richtet sich die Diagnose auf Merkmale einzelner Personen. Für die Diagnose von Schulleistung und schulleistungsrelevanten Merkmalen stehen verschiedene individualdiagnostische Verfahren (Testinstrumente) zur Verfügung, die in diesem Studienbrief exemplarisch vorgestellt werden.

Alle individualdiagnostischen Verfahren haben gemein, dass sie am Ende ein Ergebnis in Form einer Zahl liefern. Bspw. könnte das Ergebnis in einem standardisierten Lesegeschwindigkeitstest 837 lauten. Diese Zahl ist zunächst einmal nicht aussagekräftig und zwar aus zwei Gründen: Zum einen ist sie nur

durch einen Vergleich mit einer sogenannten „Bezugsnorm“ sinnvoll interpretierbar. Zum anderen macht es nur Sinn, diese Zahl überhaupt zu interpretieren, wenn der eingesetzte Lesegeschwindigkeitstest bestimmten Qualitätskriterien genügt. Es ist daher unerlässlich, dass Lehrkräfte sowohl die verschiedenen Bezugsnormen und ihre Bedeutung kennen als auch dazu in der Lage sind, diagnostische Verfahren in Bezug auf verschiedene Qualitätskriterien zu überprüfen. In Kapitel 1.2 dieses Studienbriefs können Sie sich über die verschiedenen Bezugsnormen und ihre Bedeutung informieren. Kapitel 1.3 stellt Ihnen dann die verschiedenen Qualitätskriterien für Testverfahren (die sog. „Testgütekriterien“) vor.

Testtheorie – Wozu?

Die in diesen beiden Grundlagenkapiteln zu erwerbenden testtheoretischen Kenntnisse über Bezugsnormen und Qualitätskriterien sind zum einen die Voraussetzung dafür, dass Lehrkräfte dazu in der Lage sind, aus dem Angebot existierender Testverfahren das für ihre Zwecke am besten geeignete herauszusuchen (Eine Datenbank, in der für Lehrkräfte interessante Testverfahren gelistet sind, finden Sie unter http://tests.udikom.de/). Zum anderen sind diese Kenntnisse notwendig, um die Aussagekraft individualdiagnostischer Ergebnisse, die bspw. durch den schulpsychologischen Dienst erfasst wurden, nachzuvollziehen und entsprechende pädagogische Entscheidungen ableiten zu können. Doch nicht nur die Auswahl und Interpretation bestehender individualdiagnostischer Testverfahren erfordern diese Grundlagenkenntnisse. Auch die selbstständige Entwicklung von Testverfahren – ein alltägliches Geschäft von Lehrkräften, wenn sie bspw. Klassenarbeiten konzipieren – sollte von diesen Grundkenntnissen geleitet sein. Das bedeutet nicht, dass Lehrkräfte jede ihrer Klassenarbeiten nach dem Vorbild etablierter und von wissenschaftlichen Verlagen publizierter individualdiagnostischer Testverfahren entwickeln und prüfen müssen. Dass dieser Aufwand nicht bei jeder Messung im schulischen Alltag von Lehrerinnen und Lehrern, wie z.B. bei Klassenarbeiten geleistet werden kann, ist selbstverständlich. Trotzdem sind diese testtheoretischen Kenntnisse hilfreich, um die Qualität auch von Diagnoseinstrumenten wie Klassenarbeiten zu verbessern. Wie es Lehrkräften möglich ist, das diagnostische Potenzial von Klassenarbeiten zu erhöhen, ist Gegenstand des letzten Kapitels dieses Studienbriefs.

Zusammenfassung

Individualdiagnostische Verfahren können in der pädagogischen Praxis zur Diagnose von Schulleistung und schulleistungsrelevanten Merkmalen zu verschiedenen Zeiten eingesetzt werden. Die Ergebnisse dienen als Entscheidungsgrundlage für bildungsbezogene Entscheidungen im Einzelfall. Zur Auswahl und Bewertung verschiedener individualpsychologischer Verfahren, zur Interpretation der Ergebnisse sowie für die selbstständige Entwicklung von Tests wie z.B. Klassenarbeiten sind Kenntnisse über verschiedene Bezugsnormen sowie über verschiedene Qualitätskriterien für Testverfahren notwendig. Der vorliegende Studienbrief möchte genau diese Kenntnisse und Fähigkeiten vermitteln.

1.1.1 Weiterführende Literatur

Paradies, L., Linser, H.J., & Greving, J. (2007). *Diagnostizieren, Fordern und Fördern*. Berlin: Cornelsen Verlag Scriptor.

Fisseni, H.-J. (2004). *Lehrbuch der Psychologischen Diagnostik* (3. Aufl.). Göttingen: Hogrefe.

Ingenkamp, K. (1997). *Lehrbuch der pädagogischen Diagnostik*. Weinheim: Beltz.

Schweizer, K. (2006). *Leistung und Leistungsdiagnostik*. Heidelberg: Springer.

1.2 Bezugsnormen

Fragen, die in diesem Kapitel beantwortet werden:

- *Was sind Bezugsnormen?*
- *Welchen Beitrag leisten Bezugsnormen für die Interpretation und Bewertung von Testergebnissen?*
- *Was ist bei der Nutzung der jeweiligen Bezugsnorm zu berücksichtigen?*

Rohwert

Wie in Kapitel 1.1 bereits erwähnt, haben alle Testverfahren gemeinsam, dass sie am Ende ein Ergebnis in Form einer Zahl liefern. Die interessierende Ausprägung einer Personeneigenschaft wird durch ein numerisches Testergebnis ausgedrückt. Daraus ergeben sich Aussagen wie z.B. „Die Lesegeschwindigkeit von Florian ist 837“ oder „Annas Testängstlichkeit liegt bei 3,6“. Diese Aussagen sind zunächst einmal nichtssagend. Sie enthalten zwar das numerische Testergebnis, den sogenannten „Rohwert“ (im Falle von Florian 837, bei Anna 3,6). Dieser Rohwert ist jedoch ohne weitere Informationen inhaltlich nicht interpretierbar oder bewertbar. Ist eine Lesegeschwindigkeit von 837 als gut oder schlecht zu bewerten? Schneiden Schülerinnen und Schüler, die älter als Florian sind, im Lesegeschwindigkeitstest besser als Florian ab? Ist eine Testängstlichkeit von 3,6 normal? Kann Anna mit Testsituationen jetzt besser umgehen als noch vor einem Jahr? Fragen, die für eine Interpretation und Bewertung des Testergebnisses bedeutsam sind, die aber aufgrund des Rohwerts allein nicht beantwortet werden können.

Normierung des Rohwerts

Die Interpretation und Bewertung eines Rohwerts erfolgt über einen oder mehrere Vergleiche. Dabei wird der Rohwert in Bezug gesetzt zu einer weiteren Zahl. Diese weitere Zahl dient als Norm, durch den Vergleich mit ihr wird der Rohwert normiert. Diese Norm, die in Form einer konkreten Zahl oder auch als Verteilung von Zahlen vorliegen kann, wird Bezugsnorm genannt. Sie kann als Standard oder Maßstab angesehen werden, anhand dessen das Testergebnis (und damit die Ausprägung der interessierenden Personeneigenschaft) beurteilt wird.

Bezugsnormen

Üblicherweise unterscheidet man drei verschiedene Arten von Bezugsnormen, die in Bezug auf unterschiedliche Fragestellungen der Individualdiagnostik Anwendung finden. Die *kriteriale Bezugsnorm* gibt an, wie hoch das Testergebnis einer Person, der Rohwert, mindestens sein muss, damit das Testergebnis positiv bewertet werden kann. Mit Hilfe einer *sozialen Bezugsnorm* wird das Testergebnis einer Person mit den entsprechenden Ergebnissen vergleichbarer Personen verglichen. Verwendet man eine *individuelle Bezugsnorm*, dann vergleicht man das Testergebnis einer Person mit Ergebnissen, die dieselbe Person zu früheren Zeitpunkten bereits in demselben Test erzielen konnte. Im Folgenden werden diese drei Bezugsnormen näher beleuchtet.

1.2.1 Kriteriale Bezugsnorm

Die kriteriale Bezugsnorm bezieht sich immer auf ein festes Kriterium, im Unterricht ist dies das Lehrziel, das sich ein Lehrer gesetzt hat. Unter Verwendung einer kriterialen Bezugsnorm wird bspw. überprüft, ob die Leistung eines Schülers einen bestimmten Standard erreicht oder nicht. Die kriteriale Bezugsnorm kommt auch insbesondere dann zum Einsatz, wenn der Leistungsbeurteilung eine qualifizierende oder berechtigende Funktion zukommt und somit bestimmte Standards erreicht werden müssen (Rheinberg, 2001).

Beispiel: Kriteriale Bezugsnorm

Denken wir uns einen Lehrer, der mit seiner Klasse ein Lesetraining durchgeführt hat mit dem Ziel, dass alle Schüler der Klasse einen Text flüssig und mit einer bestimmten Geschwindigkeit lesen können. Um zu überprüfen, ob er dieses Ziel erreicht hat, führt der Lehrer am Ende des Trainings einen standardisierten Lesegeschwindigkeitstest durch. Bei diesem Test wird ein Lesetext vorgegeben, den Schüler innerhalb von vier Minuten lesen sollen. Vorab definiert der Lehrer als Kriterium, dass die Schüler in der vorgegebenen Zeit mindestens 1000 Wörter verstehend lesen können. Damit dient die Zahl 1000 als kriteriale Bezugsnorm. Der Lehrer wird Schülern, die 1000 oder mehr Wörter in der gegebenen Zeit gelesen haben, eine ausreichende Lesegeschwindigkeit bescheinigen und den anderen Schülern entsprechend nicht.

1.2.2 Soziale Bezugsnorm

Bei der sozialen Bezugsnorm dienen als Vergleichswerte die Testergebnisse anderer vergleichbarer Personen. Die Beurteilung eines einzelnen Testergebnisses ist somit davon abhängig, wie hoch oder niedrig die Testergebnisse dieser vergleichbaren Personen ausfallen. Die Nutzung einer sozialen Bezugsnorm ist dann sinnvoll, wenn es bspw. Ziel der Leistungsbeurteilung ist, den Besten oder die Beste aus einer

Gruppe zu ermitteln, wie man es z.B. von sportlichen Wettkämpfen kennt. Wie Rheinberg (2001) unterstreicht, ist eine Beurteilung anhand der sozialen Bezugsnorm jedoch nicht notwendigerweise auf Selektion und Auslese der Besten beschränkt, sondern kann auch eine Grundlage für gezielte Fördermaßnahmen bieten, bspw. wenn für bestimmte Förderangebote eine nur begrenzte Anzahl von Plätzen verfügbar ist und diese Plätze denjenigen gegeben werden sollen, die sie am meisten benötigen.

Beispiel: Soziale Bezugsnorm

Nehmen wir an, ein Schüler hätte im Lesegeschwindigkeitstest innerhalb von vier Minuten 922 Wörter gelesen. Er hätte damit das Kriterium des Lehrers von 1000 Wörtern (s. Beispiel in Kap. 1.2.1) nicht erreicht und bekäme beim Anlegen einer kriterialen Bezugsnorm eine entsprechend schlechte Bewertung. Was aber, wenn eine soziale Bezugsnorm angelegt wird? In dem Fall könnte er durchaus eine sehr gute Bewertung erhalten, nämlich genau dann, wenn alle oder zumindest viele seiner Mitschüler weniger als 922 Wörter gelesen haben.

Normierung des Testverfahrens

Viele individualdiagnostische Testverfahren, die meist über Testverlage vertrieben werden, haben die Eigenschaft, dass sie „normiert" sind (s.a. Kap. 1.3). Das bedeutet, dass im Begleitheft zu den Tests sog. Normtabellen enthalten sind, die Vergleichswerte umfangreicher Stichproben präsentieren. Diese Stichproben, die häufig mehrere tausend Personen umfassen, repräsentieren bestimmte Populationen. So findet man häufig separate Normtabellen für Männer und Frauen, Normtabellen für unterschiedliche Altersklassen oder Klassenstufen, etc. Anhand dieser Normtabellen kann jeder Testanwender ein individuelles Testergebnis in Bezug auf vergleichbare Populationen einordnen, ohne dass er selbst Vergleichswerte erheben muss.

Schwächen im schulischen Kontext

Wird die soziale Bezugsnorm im schulischen Kontext angewendet, so sollten auch deren Schwachpunkte beachtet werden. Rheinberg (2001) weist auf zwei wesentliche blinde Flecken hin:

- Die soziale Bezugsnorm wird häufig auf ein klasseninternes Bezugssystem reduziert. Gilt es bspw. die Leistung eines Schülers zu bewerten, so erfolgt die Leistungsbeurteilung oftmals im Vergleich zu den entsprechenden Leistungen des jeweiligen Klassenverbands. Die Lehrkraft vergleicht innerhalb einer Klasse oder auch innerhalb einer Jahrgangsstufe. Eine einzelne Klasse oder Jahrgangsstufe ist jedoch nicht repräsentativ für eine ganze Population, und ihre durchschnittliche Leistung kann beträchtlich über oder auch unter der durchschnittlichen Leistung der Population liegen. Die Beurteilung einer einzelnen Schülerleistung hängt demnach in besonders starkem Maß davon ab, ob der Schüler in einer (verhältnismäßig) leistungsstarken oder leistungsschwachen Klasse ist. Dadurch entsteht so etwas wie der so genannte *„big fish – little pond"*-Effekt (Fischteicheffekt): Mittelstarke Schülerinnen und Schüler sind in leistungsschwachen Klassen die großen Fische im kleinen Teich, das bedeutet, dass sie – bei gleicher Leistung – besser bewertet werden als Schülerinnen und Schüler in leistungsstarken Klassen.
- Sowohl eine positive als auch eine negative Leistungsentwicklung des gesamten Klassenverbands wird bei Anwendung der sozialen Bezugsnorm ausgeblendet, da nur die Leistungsunterschiede zwischen den Schülerinnen und Schülern zählen. Das geht einher mit dem empirischen Befund, dass über 50 % der Schülerinnen und Schüler von Lehrkräften, deren Leistungsbeurteilung auf einer sozialen Bezugsnorm fußt, keine Leistungssteigerung über das Schuljahr hinweg erkennen konnten oder sogar davon ausgehen, am Ende des Schuljahres weniger zu können als am Anfang (Rheinberg, 1980).

1.2.3 Individuelle Bezugsnorm

Die individuelle Bezugsnorm ergibt sich aus den Testergebnissen einer Person, die diese zu früheren Zeitpunkten in demselben Test oder in parallelen Tests (vgl. Kap 1.3.3.1) erzielen konnte. Dadurch wird die Entwicklung eines Schülers abbildbar. Konnte sich ein Schüler in seiner Leistung steigern? Hat er etwas dazu gelernt? Hat eine Unterrichtseinheit das Interesse an dem behandelten Thema wecken oder erhöhen können? Sollen Fragen dieser Art beantwortet werden, so ist eine individuelle Bezugsnorm heranzuziehen.

Beispiel: Individuelle Bezugsnorm

Nehmen wir an, der Lehrer aus dem Beispiel in Kap. 1.2.1 hätte den Lesegeschwindigkeitstest nicht nur nach dem Lesetraining eingesetzt, sondern in einer Parallelversion auch davor. Sein Ziel sei es, die Lesegeschwindigkeit seiner Schüler zu erhöhen. Wenn seine Schüler vor dem Training im Durchschnitt 850 Wörter gelesen hätten, dann könnte der Lehrer zufrieden sein, wenn seine Schüler nach dem Training

im Durchschnitt 910 Wörter lesen könnten, da sie nach dem Training im Durchschnitt 60 Wörter mehr innerhalb der vier Minuten lesen konnten als noch vor dem Training.

Vor- und Nachteile

Das Anlegen einer individuellen Bezugsnorm kann gerade bei leistungsschwächeren Schülerinnen und Schülern enorme motivationale Vorteile bieten. Auch wenn sie in einem Test, gemäß einer kriterialen Bezugsnorm, eine eher geringe Leistung zeigen, so kann es doch sehr motivierend sein, wenn es Beachtung findet, falls diese (geringe) Leistung wenigstens besser ist als eine entsprechende Leistung zu einem früheren Zeitpunkt. Wenn man Schülerinnen und Schülern eine solche positive Entwicklung zurückmelden kann, kann dies einen enormen Effekt auf deren Motivation haben (Rheinberg, 1980).

Selbstredend hätte eine ausschließliche Beschränkung auf eine individuelle Bezugsnorm jedoch auch bizarre Folgen, bspw. wenn eine Leistungsbeurteilung eine berechtigende Funktion inne hat (z.B. Zeugnisnoten). Bei konsequenter Anwendung der individuellen Bezugsnorm würde ein Schüler, der zu Beginn eines Schuljahres eine schlechte Leistung zeigte, sich aber zum Ende des Schuljahres auf eine wenigstens durchschnittliche Leistung steigern konnte, eine bessere Zeugnisnote erhalten als ein Schüler, der von Beginn an eine durchschnittliche Leistung zeigte. Die individuelle Bezugsnorm ist also hilfreich, wenn es darum geht, die Entwicklung von Schülerinnen und Schülern abzubilden, und diese Fortschritte den Schülerinnen und Schülern auch zurückmelden zu können. Wenn der Leistungsbeurteilung jedoch eine berechtigende Funktion zukommt, dann sollte für eine solche pädagogische Entscheidung eher eine kriteriale oder eine soziale Bezugsnorm genutzt werden.

1.2.4 Bezugsnorm im Vergleich

	Kriteriale Bezugsnorm	Soziale Bezugsnorm	Individuelle Bezugsnorm
Vorteile	Bewertung unabhängig von (1) sozialen Vergleichen und von (2) der individuellen Leistungssteigerung	Ermöglicht soziale Vergleiche mit einer Bezugsgruppe	„Schwankungen im Lernverlauf werden unter individueller Bezugsnorm wie unter einem Vergrößerungsglas sichtbar gemacht“ (Rheinberg, 2001)
Nachteile	Nicht auf die Erfassung individueller Lernfortschritte ausgerichtet	Klasseninternes Bezugssystem: Big fish – little pond-Effekt Ausblendung von Leistungsschwankungen im Klassenverband	Selbstbeurteilung mittels sozialer Vergleiche nicht möglich

Tabelle 1: Vor- und Nachteile der verschiedenen Bezugsnormen

Jede der Bezugsnormen hat ihre spezifischen Vor- und Nachteile (Tabelle 1). Welche Bezugsnorm für die Interpretation eines Testwertes die richtige ist, hängt von der Art der pädagogischen Entscheidung ab.

Die individuelle Bezugsnorm ist von besonderer Bedeutung, wenn man sich mit der Entwicklung von Lernenden über einen bestimmten Zeitraum hinweg beschäftigt. So können Lernfortschritte einzelner Schülerinnen und Schüler über ein Schuljahr hinweg untersucht oder auch die Effekte pädagogischer Interventionsmaßnahmen geprüft werden. Die kriteriale Bezugsnorm kommt zum Einsatz, wenn ermittelt werden soll, inwieweit einzelne Schülerinnen und Schüler bestimmte curriculare Standards erfüllen. Die soziale Bezugsnorm wird herangezogen, um zu bilanzieren, inwiefern Schülerinnen und Schüler hinsichtlich ihrer kognitiven, affektiv-motivationalen oder beider Merkmale im Vergleich zu einer entsprechenden Bezugsgruppe mit vergleichbaren Eingangsvoraussetzungen abweichen. Auf die soziale und kriteriale Bezugsnorm wird im Studienbrief „Vergleichsarbeiten“ verstärkt eingegangen.

Die Bezugsnormen sind keinesfalls auf den pädagogisch-diagnostischen Bereich beschränkt, sondern finden sich häufig auch im Alltag wieder (Rheinberg, 2001). Nehmen wir als Beispiel die derzeit populären Castingshows, in denen der oder die beste Sängerin gekürt werden soll. Seitens der Jury oder der Moderatoren werden oft Äußerungen vorgenommen wie: „Von allen Kandidaten hat sie eindeutig die beste Performance gezeigt“ (soziale Bezugsnorm), „Er ist heute über sich hinausgewachsen. Diese Steigerung hätte in den letzten Shows niemand für möglich gehalten – eine starke Leistung" (individuelle Bezugsnorm) oder „Sie hat eine glockenklare Stimme und hat perfekt intoniert“ (kriteriale Bezugsnorm).

Nur ein Beispiel, das zeigt, dass wir in unserem Alltag, bewusst oder unbewusst, ständig auf Bezugsnormen zurückgreifen, auch wenn diese oftmals nicht die psychometrische Qualität wie bspw. die Normtabellen in standardisierten Testverfahren haben. Doch unabhängig davon ist die Verwendung einer der drei vorgestellten Bezugsnormen ein alltägliches Geschäft.

1.2.5 Weiterführende Literatur

Sacher, W. (2009). *Leistungen entwickeln, überprüfen und beurteilen. Bewährte und neue Wege für die Primar- und Sekundarstufe*. Bad Heilbrunn: Klinkhardt.

Winter, F. (2008). *Leistungsbewertung*. Hohengehren: Schneider-Verlag.

1.3 Testkonstruktion

Gegenstand dieses Kapitels ist zum einen der prinzipielle Aufbau individualdiagnostischer Testverfahren. Zudem werden Kriterien besprochen, anhand derer die Qualität von Testverfahren beurteilt werden kann. Dafür werden testtheoretische Kenntnisse vermittelt, die für ein Verständnis des Aufbaus individualdiagnostischer Verfahren sowie für ihre Bewertung unverzichtbar sind. Zu diesem Zwecke werden in diesem Kapitel folgende Fragen behandelt:

- *Was genau ist ein Test?*
- *Welche Eigenschaften hat ein Testergebnis?*
- *Was zeichnet ein qualitativ hochwertiges individualdiagnostisches Testverfahren aus?*
- *Worauf soll bei der Auswahl eines Testinstruments geachtet werden?*

1.3.1 Individualdiagnostik mit Hilfe von Tests

Was ist ein Test?

Tests, die einem Auskunft über bestimmte Eigenschaften einer Person (meist von sich selbst) versprechen, finden sich zu Hauf im Internet oder in Zeitschriften. Zehn plausibel klingende Fragen beantworten, und schon weiß man, wo die eigenen Stärken und Schwächen liegen, ob man über Sozialkompetenz verfügt, ob man ehrgeizig, einfühlsam oder auch belastbar ist, ob man gut zuhören kann oder ähnliches. Der gesunde Menschenverstand sagt einem aber bereits, dass ein „Test", der auf einer der hinteren Seiten einer TV-Zeitschrift abgedruckt ist, qualitativ wahrscheinlich nicht vergleichbar ist mit Testverfahren, die wissenschaftlich fundiert entwickelt und auf ihre Qualität hin überprüft wurden. Die Frage ist jedoch, worin genau der Unterschied in der Qualität liegt? Anhand welcher Kriterien lassen sich qualitativ hochwertige von qualitativ minderwertigen Testverfahren unterscheiden? Was sind die Voraussetzungen, die bei der Entwicklung und Überprüfung von individualdiagnostischen Tests gegeben sein müssen, damit ein qualitativ guter Test entstehen kann? Und nicht zuletzt: Was genau ist eigentlich ein Test?

Lienert (1961) betont vier Merkmale, die einen Test ausmachen, und die eine notwendige, wenn auch nicht hinreichende Voraussetzung für eine hohe Testqualität sind:

Definition: Test

„Nicht jede, zu diagnostischen Zwecken aufgestellte Untersuchung kann als Test gelten, sondern nur eine solche, die

- erstens wissenschaftlich begründet ist,
- zweitens routinemäßig – also unter Standardbedingungen mehr oder weniger handwerksmäßig durchführbar ist,
- drittens eine relative Positionsbestimmung des untersuchten Individuums innerhalb einer Gruppe oder in Bezug auf ein bestimmtes Kriterium, z.B. einem Lehrziel ermöglicht und
- viertens bestimmte empirisch – also verhaltens- und erlebnisanalytisch, phänomenologisch und nicht etwa rein begrifflich – abgrenzbare Eigenschaften, Verhaltensdispositionen, Fähigkeiten, Fertigkeiten oder Kenntnisse prüft."

Was lernen wir aus dieser Definition? Das erste angesprochene Merkmal, die wissenschaftliche Begründbarkeit, zielt darauf ab, dass ein individualdiagnostischer Test ein Personenmerkmal prüft, das sich nach wissenschaftlichen Kriterien beschreiben und begründen lässt. Anders ausgedrückt: Ein guter Test braucht eine gute theoretische wissenschaftliche Basis. Dieser Punkt wird im Laufe des Kapitels noch einmal unter dem Stichwort „Validität" besprochen. Das zweite Merkmal legt fest, dass ein Test immer in vergleichbarer Art und Weise und unter vergleichbaren Bedingungen durchgeführt, ausgewertet und interpretiert werden muss. Darauf werden wir unter den Stichworten „Objektivität" und „Reliabilität"

weiter eingehen. Das dritte Merkmal zielt auf den in Kapitel 1.2 besprochenen Umstand, dass ein Rohwert nur unter Verwendung einer Bezugsnorm inhaltlich interpretierbar ist.

Wozu testen?

Das vierte Merkmal schließlich schneidet ein Problem an, dass das Grundproblem der pädagogischen Individualdiagnostik und zugleich Anlass für die Nutzung von Testverfahren ist: Die Personenmerkmale, deren Ausprägungen man kennen möchte, um darauf aufbauend pädagogische Entscheidungen treffen zu können, sind einer direkten Empirie nicht zugänglich, d.h. sie sind nicht direkt beobachtbar (Abbildung 1). Die mathematische Fähigkeit eines Schülers oder auch seine Intelligenz ist dem Schüler nicht direkt anzusehen. Ob ein Schüler motiviert ist oder ängstlich ist, kann man nicht direkt beobachten. Was man beobachten kann sind bspw. niedergeschriebene Antworten eines Schülers in einem Mathematikleistungstest. Man kann auszählen, wie viele richtige Antworten ein Schüler in einem Intelligenztest angekreuzt hat. Man kann das mehr oder weniger ängstliche Verhalten eines Schülers beobachten oder man kann aus der Tatsache, dass ein Schüler innerhalb eines Projektes ein hohes Engagement zeigt, auf eine entsprechend hohe Motivation schließen. Das bedeutet: Wenn wir einen Test einsetzen, dann wollen wir eine Information über ein Merkmal eines Schülers, das wir nicht sehen können. Man spricht in dem Fall von einem „latenten“ Merkmal. Der Test liefert uns Information über das Schülermerkmal, in dem er uns unter möglichst standardisierten Bedingungen ein bestimmtes Verhalten (bspw. die niedergeschriebene Lösung einer Testaufgabe) eines Schülers beobachten lässt. Dieses direkt beobachtbare Verhalten bezeichnet man als „manifestes“ Merkmal. Anhand dieses Verhaltens schließen wir dann auf die Ausprägung des nicht direkt beobachtbaren, latenten Schülermerkmals. Für diesen Schluss – von dem beobachteten Verhalten auf die Ausprägung des nicht direkt beobachtbaren Personenmerkmals – ist jedoch eine Annahme notwendig, nämlich dass das beobachtete Verhalten maßgeblich von dem interessierenden latenten Personenmerkmal beeinflusst ist. Inwiefern diese notwendige Annahme berechtigt ist und unter welchen Bedingungen man davon ausgehen kann, dass diese Annahme in einer bestimmten Testsituation gültig ist, ist Gegenstand von Testtheorien, also einem Bündel theoretischer Annahmen über das Zusammenspiel nicht beobachtbarer Personenmerkmale und beobachtbarem Verhalten in Testsituationen. Die bekannteste Testtheorie, die sogenannte „Klassische Testtheorie“ wird im Folgenden weiter ausgeführt. Eine weitere Testtheorie, die insbesondere im Rahmen sogenannter „large-scale assessments“ Anwendung findet, können Sie ausführlich in den Studienbriefen „Vergleichsarbeiten“ und „Bildungsmonitoring“ kennen lernen.

Latentes und manifestes Merkmal

Testtheorie

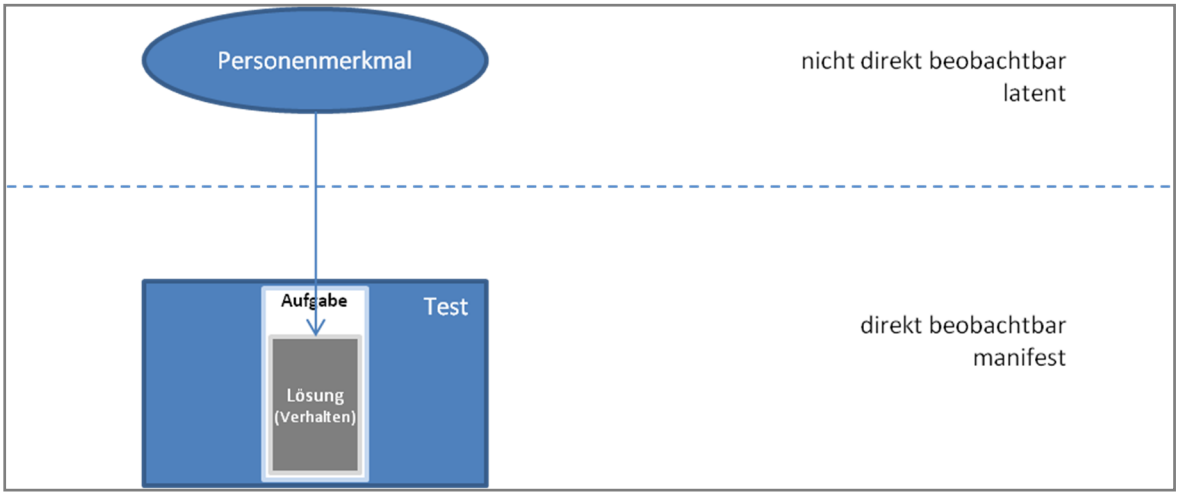


Abbildung 1: Annahme über den Zusammenhang zwischen dem Personenmerkmal und dem Lösen einer Testaufgabe

1.3.2 Das Testergebnis

Angenommen, wir wollten überprüfen, ob ein Schüler dazu in der Lage ist, einen für seine Altersgruppe angemessenen Text verstehend zu lesen. Dafür würden wir ihm einen entsprechenden Text geben, mit der Bitte ihn innerhalb einer bestimmten Zeit zu bearbeiten und zu versuchen, den Textinhalt zu verstehen. Um das Textverständnis zu überprüfen, würden wir dem Schüler nach dem Lesen drei Fragen zum Text stellen, die er kurz schriftlich beantworten soll (Abbildung 2). Das zu überprüfende latente Personenmerkmal wäre in diesem Beispiel das Textverständnis (im Sinne der Fähigkeit, einen Text verstehend zu lesen). Der Test bestünde aus dem Text und den drei Fragen. Das beobachtbare manifeste Verhalten wären die drei Antworten auf die drei Fragen.

Unsere grundlegende Annahme wäre, dass das Textverständnis des Schülers dafür verantwortlich ist, ob wir von ihm gute und richtige Antworten geliefert bekommen oder nicht. Diese sehr einfache, aus genau

einer Annahme bestehende Testtheorie ist jedoch leider nicht haltbar. Die Annahme, dass die Antwort in einem Test ausschließlich von der Fähigkeit einer Person abhängt, ignoriert, dass es noch viele weitere Einflüsse auf das in einem Test gezeigte Verhalten geben kann. Bspw. könnte der Schüler zwar über ein hohes Textverständnis verfügen, jedoch wenig motiviert sein, sich testen zu lassen, weshalb seine Antworten nicht so gut ausfallen wie möglich. Oder aber der Schüler könnte bei einer Frage zwar unsicher gewesen sein, aber trotzdem mit etwas Glück eine gute Antwort aufgeschrieben haben. Evtl. war der Schüler beim Lesen eines Textabschnitts durch einen Mitschüler abgelenkt, so dass er, bei eigentlich gutem Textverständnis, für eine diesen Abschnitt betreffende Frage keine gute Antwort abliefert. Die Anzahl verschiedener möglicher Einflüsse auf das in einem Test gezeigte Verhalten ist nahezu unendlich groß. Aus diesem Grund ist es auch müßig zu versuchen, jeden einzelnen Einfluss genau zu benennen, zumal die Hoffnung besteht, dass jeder einzelne dieser Einflüsse verschwindend gering ist. Deshalb werden diese Einflüsse üblicherweise unter dem Begriff „Fehler“ einfach zusammengefasst. Welche Eigenschaften dieser Fehler hat, wie er das Testverhalten beeinflusst, wie man seine Größe bestimmen kann und wie man Tests konstruieren kann, bei denen der Einfluss des Fehlers auf das Testverhalten möglichst gering ist, das ist Gegenstand der sogenannten „Klassischen Testtheorie“.

Zufällige Einfüsse = Fehler

Die Klassische Testtheorie besteht aus einem Satz von Aussagen (Axiome), der den theoretischen Hintergrund für die meisten individualdiagnostischen Verfahren bildet. Kennzeichnend für die Klassische Testtheorie ist zum einen die Annahme, dass das Testergebnis zwar maßgeblich von der Ausprägung des latenten Personenmerkmals abhängt, dass es aber zusätzlich noch einen „Messfehler“ beinhaltet. Da dieses eine der zwei zentralen Annahmen der Klassischen Testtheorie ist, wird sie oftmals auch als Fehlertheorie bezeichnet.

Messwert = wahrer Wert + Messfehler

Der Messfehler (kurz: Fehler) repräsentiert den zufälligen und unsystematischen Anteil des Testergebnisses, der nicht auf das latente Personenmerkmal, sondern auf situative Zufälligkeiten zurückgeführt werden muss (Gröschke, 2005). Gemäß der Klassischen Testtheorie setzt sich damit ein Messwert, sprich das Testergebnis, immer aus zwei Anteilen additiv zusammen: dem wahren Wert, der auf die Ausprägung des Personenmerkmals zurückgeführt werden kann, und dem Messfehler.

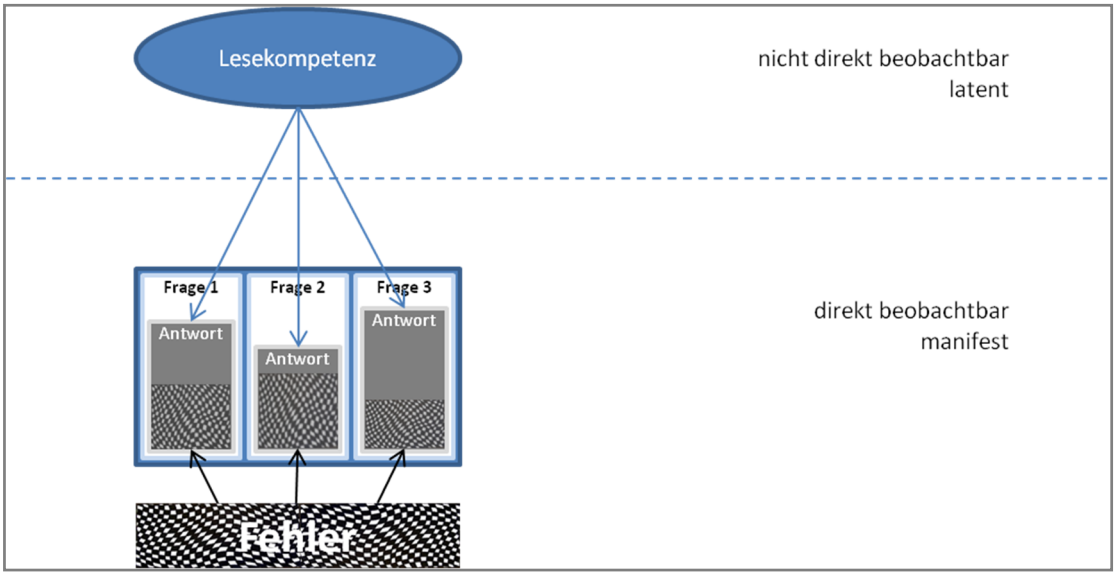


Abbildung 2: Einfluss des Personenmerkmals sowie zufälliger Gegebenheiten (Fehler) auf das Verhalten in Tests

Fehler = Zufall

Die zweite zentrale Annahme der Klassischen Testtheorie ist, dass der Messfehler, wie bereits erwähnt, rein zufällig ist. Aus dieser Annahme ergibt sich, dass der Messfehler unabhängig ist von jeglichen Gegebenheiten. Wie hoch der Anteil eines Messergebnisses ist, das auf zufällige Umstände zurückgeführt werden muss, ist bspw. unabhängig von der Ausprägung des eigentlich interessierenden Personenmerkmals, dem latenten Merkmal. Er ist unabhängig von der Ausprägung weiterer Merkmale der Person, und er ist unabhängig von Messfehlern, die bei früheren Einsätzen des Testinstruments zu verzeichnen waren. Diese (durch die Klassische Testtheorie postulierte) Eigenschaft des Fehlers ist von zentraler Bedeutung. Aus ihr lässt sich herleiten, wie der Fehleranteil an einem Testergebnis reduziert werden kann. Dieses werden wir im nächsten Kapitel „Indikatorbildung“ besprechen. Zum anderen macht man sich die Zufälligkeit des Fehlers zu Nutze, wenn für ein individualdiagnostisches Testverfahren überprüft werden soll, wie groß der Messfehleranteil am Testergebnis ist. Wir werden im Kapitel „Reliabilität“ auf diesen Punkt zurückkommen.

1.3.3 Indikatorbildung

Üblicherweise besteht ein Test nicht nur aus einer Aufgabe, sondern aus mehreren bis vielen Testaufgaben. Bspw. besteht unser Leseverständnistest nicht nur aus einer, sondern aus drei Aufgaben. Für jede Aufgabe können wir beobachten, ob der Schüler sie korrekt bearbeitet oder nicht. Wurde eine Aufgabe gelöst, bekommt der Schüler dafür einen oder mehrere Punkte (=Messwerte), ansonsten eben nicht. Damit bekommen wir für jeden Schüler so viele Zahlen (Messwerte) wie der Test Aufgaben hat. Da wir die Ausprägung des Personenmerkmals – in unserem Fall das Ausmaß der Fähigkeit, einen Text verstehend zu lesen – aber durch nur eine Zahl ausdrücken möchten, berechnen wir üblicherweise die Summe oder den Mittelwert der Punkte über alle Testaufgaben hinweg. Diese Summe bzw. dieser Mittelwert ist dann das Testergebnis. Es dient als Indikator für die Ausprägung des latenten Personenmerkmals. Die Aufgaben, deren Punkte durch eine Summe oder einen Mittelwert zusammengefasst werden, bilden eine sogenannte „Skala". Eine Skala ist die Messlatte, mit deren Hilfe ein Messwert bestimmt wird. Sie erstreckt sich von dem kleinsten Wert, den die Summe oder der Mittelwert annehmen kann, bis hin zu dem entsprechenden größten Wert. Bekäme bspw. in unserem Leseverständnistest ein Schüler bei jeder Aufgabe für ihre korrekte Bearbeitung genau einen Punkt (und ansonsten null Punkte), dann erhielten wir bei drei Aufgaben durch Summenbildung eine Skala mit Werten von 0 bis 3.

Skala

Die Summenbildung hat jedoch nicht nur das Ziel, mit nur möglichst einem Indikator die Ausprägung des Personenmerkmals einschätzen zu können. Viel wichtiger ist, dass auf diese Art der Fehleranteil am Testergebnis reduziert wird. Und der Grund dafür liegt in der Zufälligkeit des Fehlers: Die Antwort jeder Aufgabe, sprich jeder Messwert, enthält einen Fehler. Dabei ist es aber zum einen vollkommen zufällig, wie groß dieser Fehler ist. Zum anderen ist es ebenfalls vollkommen zufällig, ob der Fehler dazu führt, dass der Messwert den wahren Wert überschätzt oder unterschätzt (ob also der Fehler zum wahren Wert addiert oder von ihm subtrahiert werden muss). Bildet man jetzt die Summe über mehrere Messwerte, die jeweils entweder einen „überschätzenden" oder einen „unterschätzenden" Fehler enthalten, dann subtrahiert man automatisch von der Summe der überschätzenden Fehler die Summe der unterschätzenden Fehler. Dadurch reduziert sich automatisch der Anteil der Fehler an der Summe der Messwerte, sprich am Testergebnis. Im Idealfall ist die Summe der überschätzenden Fehler gleich der Summe der unterschätzenden Fehler. In dem Fall würden sich beide Fehlersummen gegenseitig aufheben und der Anteil des Fehlers am Testergebnis wäre Null.

Reduktion des Fehlers

Dieser Idealfall ist natürlich mehr als selten. Aber man kann nahe an ihn herankommen. Der Trick ist, die Anzahl der Testaufgaben zu erhöhen. Je mehr Aufgaben ein Test enthält, desto höher ist die Wahrscheinlichkeit, dass die Summe der überschätzenden Fehler gleich der Summe der unterschätzenden Fehler wird. Anders ausgedrückt, je mehr Aufgaben ein Test enthält, desto höher ist die Wahrscheinlichkeit, für einen Messwert mit einem Fehleranteil von +x einen Messwert mit einem Fehleranteil von –x zu finden. Summiert man beide Fehleranteile, so ergibt sich Null.

Viele Aufgaben = wenig Fehler

1.3.4 Kriterien der Qualität von Tests

Nachdem wir uns erarbeitet haben, was ein individualdiagnostischer Test ist und welche Eigenschaften ein Ergebnis hat, wenden wir uns nun der Frage zu, woran man die Qualität eines Tests erkennen kann bzw. auf welche Art und Weise man diese gewährleisten und überprüfen kann. Bei der Bewertung der Qualität von Tests orientiert man sich zunächst einmal hauptsächlich an drei Testgütekriterien, nämlich an der Reliabilität, der Validität sowie der Objektivität. Wir werden diese im Folgenden genauer besprechen. Neben diesen sogenannten „Hauptgütekriterien" gibt es jedoch auch eine Menge weiterer sogenannter „Nebengütekriterien". Zu den wichtigsten zählt dabei sicherlich die Normierung eines Tests. Normierungen von Tests werden Gegenstand des nachfolgenden Kapitels 1.3.4 sein.

1.3.4.1 Reliabilität

Wie bereits erwähnt, nimmt die Klassische Testtheorie an, dass ein Testergebnis zwar ein guter Indikator für die latente Merkmalsausprägung sein kann, dass das Ergebnis jedoch durch den Messfehler verzerrt wird. Wie hoch der Fehleranteil am Testergebnis ist, das ist die Frage nach der sogenannten Reliabilität des Tests. Die Reliabilität eines Tests spiegelt die Zuverlässigkeit oder die Genauigkeit eines Tests wider. Sie ist umso höher, je geringer der Fehleranteil am Testergebnis ist.

Definition: Reliabilität

Die Reliabilität eines Tests kennzeichnet den Grad der Genauigkeit, mit dem das geprüfte Merkmal gemessen wird (Bortz, 2005).

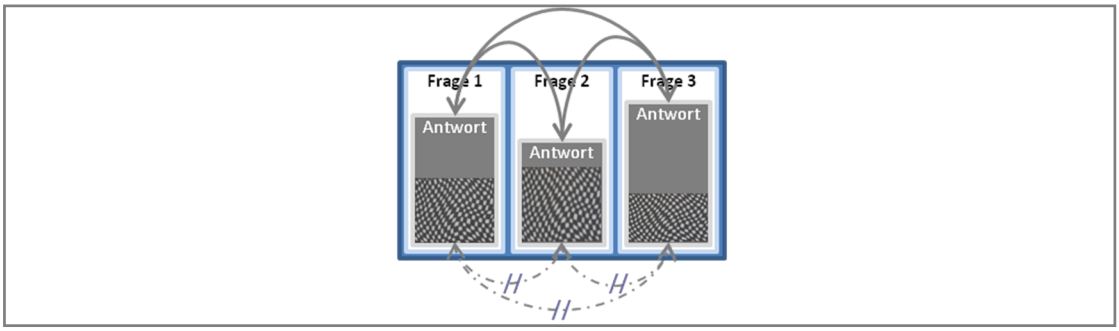


Abbildung 3: Reliabilität - Zusammenhänge der wahren Werte und Unabhängigkeiten der Messfehler

Es stellt sich jedoch die Frage, wie die Reliabilität, d.h. die Messgenauigkeit eines Tests ermittelt werden kann. Die Annahmen der Klassischen Testtheorie bieten die Grundlage für eine Antwort. Eine der beiden zentralen Annahmen der Klassischen Testtheorie ist, dass die Ausprägung des Messfehlers rein zufällig ist. Das bedeutet, dass der Messfehler von allen Gegebenheiten unabhängig ist, sprich der Messfehler steht in keinerlei Zusammenhang mit irgendetwas anderem (Kap. 1.3.2). Anders verhält es sich jedoch mit dem Anteil am Messwert, der durch die Ausprägung des zu messenden latenten Personenmerkmals bedingt ist, also dem wahren Wert. Der wahre Wert ist abhängig von der Ausprägung des Personenmerkmals, sprich der wahre Wert steht in systematischem Zusammenhang mit dem Personenmerkmal. Dadurch steht dieser wahre Wert jedoch auch in systematischem Zusammenhang mit den wahren Werten von Messergebnissen, die beim Testen desselben Personenmerkmals mit verschiedenen Aufgaben oder aber bei einer wiederholten Durchführung desselben Tests erzielt wurden.

Die wahren Werte stehen in starkem Zusammenhang (sprich: sie sind korreliert), die Messfehler stehen in keinem Zusammenhang (Abbildung 3). Diese zentrale Annahme der Klassischen Testtheorie kann man sich zu Nutze machen, um die Reliabilität eines Testverfahrens zu bestimmen. Das Grundprinzip ist dabei immer dasselbe: Es wird überprüft, wie stark der Zusammenhang zwischen wiederholten oder verschiedenen Messungen desselben Personenmerkmals ist. Je höher dieser Zusammenhang ist, desto höher muss der Anteil der wahren Werte an den Messergebnissen sein, da ausschließlich die wahren Werte in Zusammenhang zueinander stehen. Ein hoher Fehleranteil an den Messwerten würde dazu führen, dass die Messwerte in keinem Zusammenhang miteinander stehen. Die Messwerte wären unkorreliert.

Berechnung der Reliabilität

Um das Ausmaß an Genauigkeit eines Tests ausdrücken zu können, wird meist ein sogenannter Korrelationskoeffizient berechnet. Dieser drückt durch eine Zahl die Stärke des Zusammenhangs zweier Variablen aus. Bspw. kann ein Korrelationskoeffizient berechnet werden für den Zusammenhang zwischen den Punkten, die in zwei Testantworten erzielt wurden. Wurde derselbe Test mit denselben Personen zu zwei Zeitpunkten wiederholt durchgeführt, kann der Zusammenhang zwischen den Testleistungen zu den beiden Testzeitpunkten durch die Berechnung eines Korrelationskoeffizienten ausgedrückt werden.

Definition: Korrelationskoeffizient

Der Zusammenhang zwischen zwei Testergebnissen wird als Korrelation bezeichnet. Die Stärke des Zusammenhangs wird durch den Korrelationskoeffizienten ausgedrückt, der wiederum durch den Buchstaben r abgekürzt wird (z.B. $r = 0{,}65$). Dieser Korrelationskoeffizient r kann einen Wert zwischen -1 und +1 annehmen. Werte, die gegen +1 oder auch gegen -1 streben, zeigen einen starken Zusammenhang an, während Werte nahe Null für einen schwachen bis gar keinen Zusammenhang stehen. Das Vorzeichen enthält die Information über die Art des Zusammenhangs: Ein positives Vorzeichen steht für einen gleich gerichteten Zusammenhang (Je größer…, desto größer…), ein negatives Vorzeichen steht für einen entgegengesetzt gerichteten Zusammenhang (Je größer…, desto kleiner…).

Es gibt verschiedene Formen, die Reliabilität eines Tests zu bestimmen. Sie werden auf den folgenden Seiten kurz dargestellt. Das Grundprinzip ist dabei aber immer dasselbe. Über die Berechnung eines Korrelationskoeffizienten (oder aber einer bestimmten Abwandlung des Korrelationskoeffizienten) wird die Genauigkeit des Tests durch eine Zahl ausgedrückt und so bewertbar gemacht.

Test-Retest-Reliabilität (Stabilität)

Für die Bestimmung der Test-Retest-Reliabilität wird derselbe Test mit einer Gruppe von Personen in einem größeren zeitlichen Abstand zweimal durchgeführt (Abbildung 4). Man spricht daher auch von

der „Testwiederholungsmethode“. Berechnet wird dann der Korrelationskoeffizient (kurz: die Korrelation) für den Zusammenhang zwischen den Ergebnissen beider Testzeitpunkte. Dabei entsteht ein hoher Zusammenhang, wenn die Personen, die beim ersten Testzeitpunkt ein hohes Ergebnis erzielen konnten, dieses auch beim zweiten Testzeitpunkt erzielen, und wenn in gleichem Maße Personen, die beim ersten Mal eher niedrige Ergebnisse erreichten, auch beim zweiten Mal eher niedrige Testwerte erreichen.

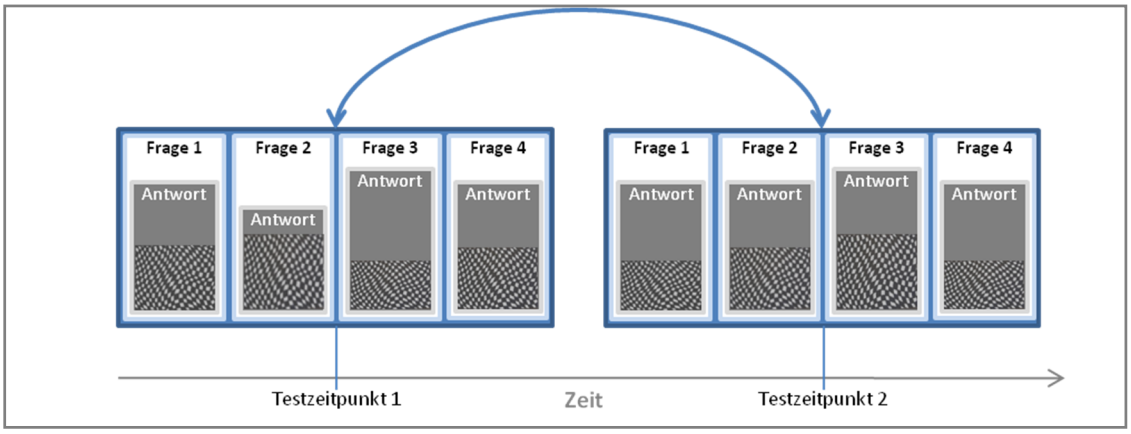


Abbildung 4: Test-Retest-Reliabilität

Beispiel: Test-Retest-Reliabilität

Angenommen, unserer Lehrer aus den Beispielen in Kapitel 1.2 hätte Zweifel an der Reliabilität seines Lesegeschwindigkeitstests. Um diese zu überprüfen, würde er den Test in einer seiner Schulklassen in einem zeitlichen Abstand von drei Wochen zweimal durchführen. Für jeden Testzeitpunkt könnte er dann zunächst seine Schüler nach der erreichten Punktzahl im Test sortieren, sie also in eine Rangreihe bringen. Wenn die Rangreihen für die beiden Testzeitpunkte sehr stark korrespondierten, dann spräche das für eine gute Reliabilität des eingesetzten Lesegeschwindigkeitstests. Statt dem Vergleichen von Rangreihen könnte der Lehrer aber natürlich auch die Korrelation berechnen (Wie das mit Hilfe von Microsoft Excel recht einfach geht, ist Gegenstand von Kapitel 1.5). Wenn die beiden Rangreihen sehr ähnlich oder identisch sind, dann strebt der Korrelationskoeffizient gegen $r = +1$.

Aus dem Beispiel werden wenigstens zwei Bedingungen deutlich, die gegeben sein müssen, um die Testhalbierungsmethode einsetzen zu können. Zum einen sollte man von dem Personenmerkmal, das gemessen werden soll – im Beispiel also die Lesegeschwindigkeit –, annehmen können, dass es sich nicht über die drei Wochen hinweg verändert. Zum anderen muss die Zeitspanne zwischen den beiden Testzeitpunkten groß genug sein, um ausschließen zu können, dass das Durchführen des ersten Tests Einfluss auf die Ergebnisse des zweiten Tests hat. Dies könnte bspw. durch Lerneffekte der Fall sein.

Paralleltest-Reliabilität (Äquivalenz)

Die Methode kommt dann zum Einsatz, wenn aus praktischen Gründen zwei äquivalente Testversionen erforderlich sind, bspw. um zu verhindern, dass die Testpersonen von ihren Nachbarn die Lösungen abschreiben (Abbildung 5). Um die Paralleltest-Reliabilität zu ermitteln, werden zwei Versionen eines Tests entwickelt, die beide gleichermaßen auf die Erfassung desselben Personenmerkmals abzielen. Beide Tests werden mit denselben Testpersonen in kurzem zeitlichem Abstand innerhalb einer Sitzung durchgeführt. Im Anschluss wird die Korrelation der beiden Testversionen berechnet. Diese Methode ist bei der Entwicklung eines Tests natürlich recht aufwändig, da man streng genommen nicht nur einen, sondern zwei Tests entwickeln muss. Sie bietet sich entsprechend nur an, wenn man aus praktischen Gründen (s.o.) sowieso zwei Testversionen braucht.

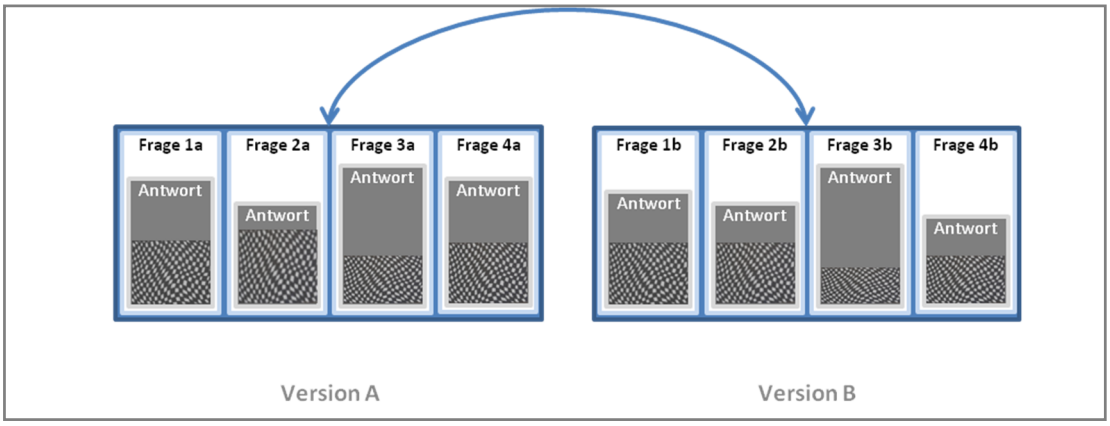


Abbildung 5: Paralleltest-Reliabilität

Testhalbierungs-Reliabilität

Im Gegensatz zur Paralleltest-Methode ist die Testhalbierungsmethode mit geringem Zusatzaufwand für die Testentwicklung verbunden. Der Test wird dafür nur einmal mit einer Gruppe von Personen durchgeführt. Im Anschluss werden die Aufgaben des Tests in zwei Gruppen aufgeteilt (z.B. „split-half-Methode“: die erste Hälfte der Aufgaben versus die zweite Hälfte der Aufgaben oder „odds-even-Methode“: die Aufgaben mit einer geraden Aufgabennummer versus die Aufgaben mit einer ungeraden Aufgabennummer). Die beiden Aufgabengruppen werden dann wie zwei separate Tests behandelt, und es wird für jede Aufgabengruppe das Testergebnis berechnet. Danach berechnet man, ähnlich wie bei der Paralleltest-Methode, die Korrelation dieser beiden Testergebnisse (Abbildung 6).

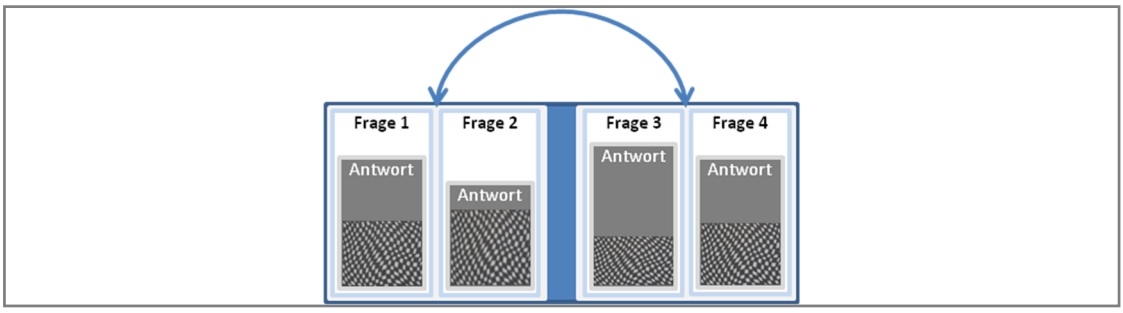


Abbildung 6: Testhalbierungs-Reliabilität

Interne Konsistenz

Die sogenannte „interne Konsistenz“ ist ebenfalls ein Schätzer für die Reliabilität eines Tests. Der Gedanke dahinter ist, dass die Antworten einer Person auf Aufgaben, die alle dasselbe Personenmerkmal abbilden sollen, von ähnlicher Qualität sein sollten. Entsprechend sollten alle Aufgaben eines Tests in engen Zusammenhängen zueinander stehen, was als interne Konsistenz des Tests bezeichnet wird. Bestimmt wird die interne Konsistenz, indem die Korrelationen einer jeden Aufgabe mit allen anderen Aufgaben berechnet (Abbildung 7) und unter Berücksichtigung bestimmter Gewichte zu einem Mittelwert zusammengefasst werden. Der resultierende Koeffizient nennt sich „Cronbachs Alpha“ (Cronbach, 1951). Auch er kann theoretisch wie ein Korrelationskoeffizient Werte zwischen -1 und +1 annehmen. Negative Werte sind jedoch äußerst selten. Von einer hohen internen Konsistenz geht man aus, wenn Cronbachs $\alpha \geq +0{,}8$ ist.

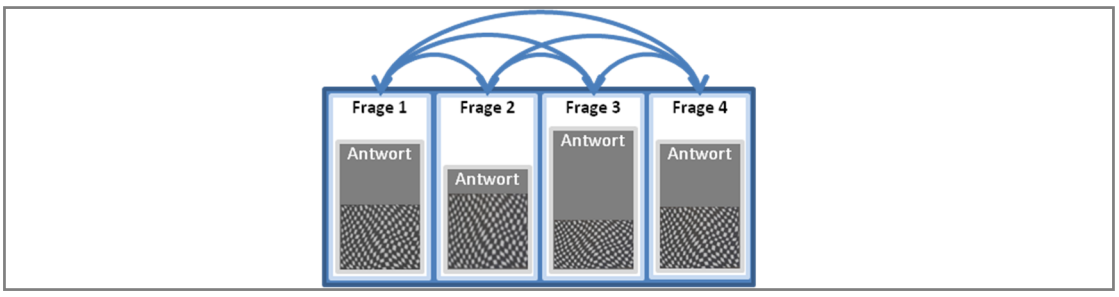


Abbildung 7: Interne Konsistenz

1.3.4.2 Validität

Die Validität eines Tests ist das Merkmal, das in der Lienertschen Definition eines Tests (Kap. 1.3.1) unter dem Stichwort „wissenschaftliche Begründbarkeit“ als erstes angesprochen wurde, was die Bedeutung dieses Testgütekriteriums unterstreicht. Sie betrifft die Frage, ob die Antworten in einem Test tatsächlich maßgeblich von dem vermuteten Personenmerkmal abhängen oder aber von einem ganz anderen. Sind die Antworten im Leseverständnistest tatsächlich von der Lesekompetenz der Schülerinnen und Schüler geprägt oder gibt es zusätzlich oder alternativ dazu weitere Personenmerkmale, die die Antworten im Lesekompetenztest systematisch beeinflussen? Welchen Einfluss hat bspw. die Motivation eines Schülers, in diesem Test eine möglichst gute Leistung zu zeigen (Abbildung 8)? Oder nehmen wir einen Test zur Erfassung mathematischer Kenntnisse und Fähigkeiten, der Textaufgaben enthält. Sind die Antworten auf diese Aufgaben wirklich ausschließlich durch die mathematischen Kenntnisse und Fähigkeiten des Schülers bedingt oder werden sie zusätzlich von dessen Lesekompetenz beeinflusst? Und falls ja, wie stark ist dann dieser zusätzliche Einfluss der Lesekompetenz? Ist er so gering, dass man ihn vernachlässigen kann oder so stark, dass man nicht mehr davon ausgehen kann, dass die manifesten Testantworten gute Indikatoren für die latenten mathematischen Fähigkeiten sind?

Definition: Validität

Die Validität (Gültigkeit) eines Tests gibt an, wie gut der Test in der Lage ist, genau das zu messen, was er zu messen vorgibt (Bortz, 2005).

Bedeutung der Validität

Die grundlegende Voraussetzung für die Validität eines Tests ist die klare theoretische Definition des Personenmerkmals, das getestet werden soll. In Bezug auf den Lesekompetenztest z.B. müssen wir uns fragen, was genau wir unter Lesekompetenz verstehen? Ist Lesekompetenz gleich Leseverständnis? Und was genau ist Leseverständnis? Gehört zur Lesekompetenz neben dem Leseverständnis auch die Lesegeschwindigkeit? Sich solche und andere Fragen zu stellen ist insbesondere für zwei Punkte wichtig. Zum einen ist die Beantwortung dieser Fragen eine notwendige Voraussetzung, wenn ein individualdiagnostisches Testverfahren neu entwickelt werden soll. Inwieweit im schulischen Kontext dies nötig ist und wie in einem solchen Fall diese Fragen beantwortet werden können, ist Gegenstand des Kapitels 1.5. Zum anderen müssen diese Fragen beantwortet werden, wenn aus der Fülle bereits existierender Testverfahren eines ausgesucht werden soll, das einem genau die Informationen liefert, die für eine pädagogische Entscheidung benötigt werden (Für die Suche und Auswahl individualdiagnostischer Testverfahren im pädagogischen Kontext siehe die UDiKom-Testdatenbank, die im Internet unter http://tests.udikom.de/ frei verfügbar ist).

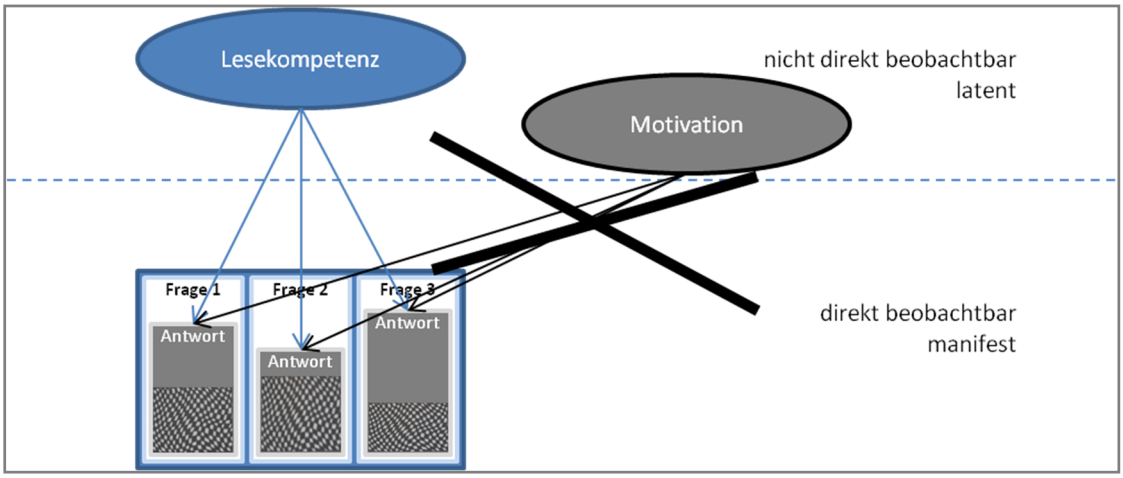


Abbildung 8: Validität

Bedeutung theoretischer Modelle

Doch wie kann man diese Fragen beantworten? Hier kommt die wissenschaftliche Begründbarkeit ins Spiel. Die entsprechenden wissenschaftlichen Disziplinen wie die Erziehungswissenschaft oder die Pädagogische Psychologie liefern Theorien und Modelle, mit deren Hilfe die latenten Personenmerkmale (vgl. Kap. 1.3.1) beschrieben werden können. Für eine gute pädagogische Individualdiagnostik ist es unerlässlich, sich vorab auf der Basis solcher wissenschaftlich begründeter theoretischer Modelle ein Bild von dem nicht direkt beobachtbaren Personenmerkmal zu machen. Dies ist die notwendige Grundlage, um entweder dazu passende Testverfahren auszuwählen oder aber selbst zu entwickeln.

Dimensionen theoretischer Modelle

Die unterschiedlichen Theorien und Modelle lassen sich anhand verschiedener Merkmale klassifizieren, wobei wir uns im Rahmen dieses Studienbriefs auf die Besprechung eines Merkmals beschränken wollen, nämlich das Merkmal der Dimensionalität. Die Dimensionalität betrifft die Frage, wie komplex ein Personenmerkmal ist bzw. wie komplex das beschreibende Modell sein muss. Ist bspw. die Lesekompetenz ein sehr wenig komplexes Personenmerkmal oder muss man innerhalb dieser Kompetenzen wiederum verschiedene Teilkompetenzen unterscheiden wie z.B. Leseverständnis und Lesegeschwindigkeit? Je mehr Teilkompetenzen oder Facetten durch ein Modell beschrieben werden, um das interessierende Personenmerkmal umfassend abbilden zu können, desto mehr Dimensionen hat das Modell. Für das entsprechende Testverfahren bedeutet dies, dass es nicht nur aus einer Skala besteht, sondern aus genau so vielen Skalen (vgl. Kap. 1.3.3) wie das Modell Dimensionen hat. Diese Skalen sind Untertests, deren Qualität genauso überprüft und bewertet werden muss, wie die entsprechende Qualität eines eindimensionalen Tests.

Unabhängig davon, ob ein Modell und damit ein Test eindimensional oder mehrdimensional ist, muss gewährleistet sein, dass der Test valide ist. Im Rahmen des Studienbriefs wollen wir drei gängige Arten der Validierung besprechen. Die erste ist die sogenannte „*Inhaltsvalidität*". In diesem Fall geht es weniger um eine Überprüfung der Validität, sondern stärker um das Vorgehen bei der Entwicklung von Testverfahren, die die Validität gewährleisten sollen. Die weiteren beiden Validierungsarten, die „*Kriteriumsvalidität*" sowie die „*Konstruktvalidität*", sind dann jedoch Arten der empirischen Überprüfung der Validität. Hierfür müssen der zu validierende Test und andere Tests und Erfassungsmethoden bei einer Stichprobe von Personen eingesetzt werden, was einen gewissen Aufwand bedeutet.

Inhaltsvalidität

Die Inhaltsvalidität, auch „Kontentvalidität" genannt, ist besonders bei Testverfahren relevant, die sich einem bestimmten Fach, einem Themenbereich oder auch einer bestimmten Unterrichtseinheit zuordnen lassen. Sie gibt an, wie gut die Aufgaben eines Tests den zu testenden Inhaltsbereich repräsentieren. Hierbei geht es insbesondere um die Frage, ob alle relevanten Aspekte eines Inhaltsbereichs durch die Aufgaben umfassend abgebildet werden. Die Inhaltsvalidität wird meist durch Expertenbefragungen oder ähnliche Verfahren eingeschätzt oder aber durch die Art der Aufgabenkonstruktion und -auswahl sichergestellt (vgl. Kap. 1.5).

Definition: Inhaltsvalidität

„Inhaltsvalidität ist gegeben, wenn der Inhalt der Testitems das zu messende Konstrukt in seinen wichtigsten Aspekten erschöpfend erfasst […]. Hieraus folgt jedoch, dass die Grundgesamtheit der Testitems, die potentiell für die Operationalisierung eines Items in Frage kommen, sehr genau definiert werden muss. Die Inhaltsvalidität eines Tests ist umso höher, je besser die Testitems diese Grundgesamtheit repräsentieren" (Bortz & Döring, 2006, S. 200).

Kriteriumsvalidität

Bei der Kriteriumsvalidität wird überprüft, inwiefern eine Testleistung (und damit das interessierende latente Personenmerkmal) in Zusammenhang mit einem anderen direkt beobachtbaren manifesten Personenmerkmal steht. Es geht also, wie bereits bei der Bestimmung der Reliabilität, um die Stärke eines Zusammenhangs, die sich empirisch ermitteln und mittels eines Korrelationskoeffizienten numerisch angeben lässt. Im Gegensatz zur Reliabilitätsbestimmung wird jetzt jedoch nicht der Zusammenhang zwischen Aufgaben (bzw. Antworten) desselben oder wiederholt durchgeführten Tests überprüft, sondern der Zusammenhang zwischen dem Testergebnis und einem Personenmerkmal, das nicht durch den eigentlichen Test erfasst wird (Abbildung 9). Zudem handelt es sich bei diesem Personenmerkmal um ein manifestes, sprich direkt beobachtbares Merkmal.

Definition: Kriteriumsvalidität

Ein Test weist Kriteriumsvalidität auf, wenn vom Verhalten der Testperson innerhalb der Testsituation erfolgreich auf ein Kriterium, nämlich auf ein Verhalten außerhalb der Testsituation, geschlossen werden kann (Moosbrugger & Kevala, 2008).

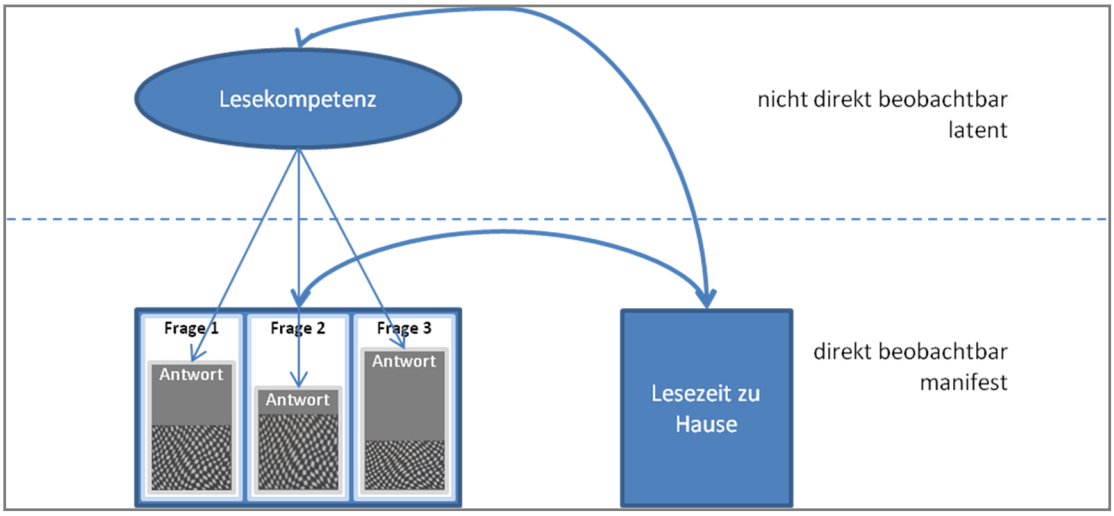


Abbildung 9: Kriteriumsvalidität

Beispiel: Kriteriumsvalidität

Es ist bekannt, dass Schülerinnen und Schüler, die zu Hause gerne und viel lesen, über eine höhere Lesekompetenz verfügen als Schülerinnen und Schüler, die nur ungern lesen. Wenn man die Kriteriumsvalidität eines Lesefähigkeitstests überprüfen wollte, könnte man entsprechend untersuchen, ob sich empirisch eine Korrelation zwischen dem Testergebnis und der durchschnittlichen Lesezeit zu Hause nachweisen lässt (Die Zeit, die jemand mit Lesen verbringt, ist direkt beobachtbar). Je höher der Korrelationskoeffizient, desto höher wäre die Kriteriumsvalidität einzuschätzen.

Konstruktvalidität

Anstelle eines direkt beobachtbaren Außenkriteriums, wie bei der Kriteriumsvalidität, werden zur Überprüfung der Konstruktvalidität latente, nicht direkt beobachtbare Personenmerkmale herangezogen (Abbildung 10). Ihre Bezeichnung erhält die Konstruktvalidität daher, dass latente, nicht direkt beobachtbare Personenmerkmale, die die zentrale Rolle bei der Konstruktvalidierung spielen, üblicherweise auch als „Konstrukt" bezeichnet werden. Auf der Basis entsprechender theoretischer Modelle sowie wissenschaftlicher Erkenntnisse über das Zusammenspiel der damit beschriebenen Personenmerkmale werden Hypothesen über Zusammenhänge zwischen dem zu erfassenden und anderen latenten Merkmalen formuliert und überprüft. Dabei können sowohl Vermutungen über starke Zusammenhänge als auch Vermutungen über Unabhängigkeiten (kein Zusammenhang) aus der Literatur abgeleitet und überprüft werden. Geht man davon aus, dass zwischen zwei latenten Personenmerkmalen ein Zusammenhang besteht, spricht man von *konvergenter Validität*. Ist kein Zusammenhang zu vermuten, wird von *diskriminanter Validität* gesprochen.

Konstrukt = latentes Merkmal

Konvergente und diskriminante Validität

Definition: Konstruktvalidität

Ein Test ist konstruktvalide, wenn aus dem zu messenden Zielkonstrukt Hypothesen ableitbar sind, die anhand der Testwerte bestätigt werden können (Bortz & Döring, 2006, S. 201).

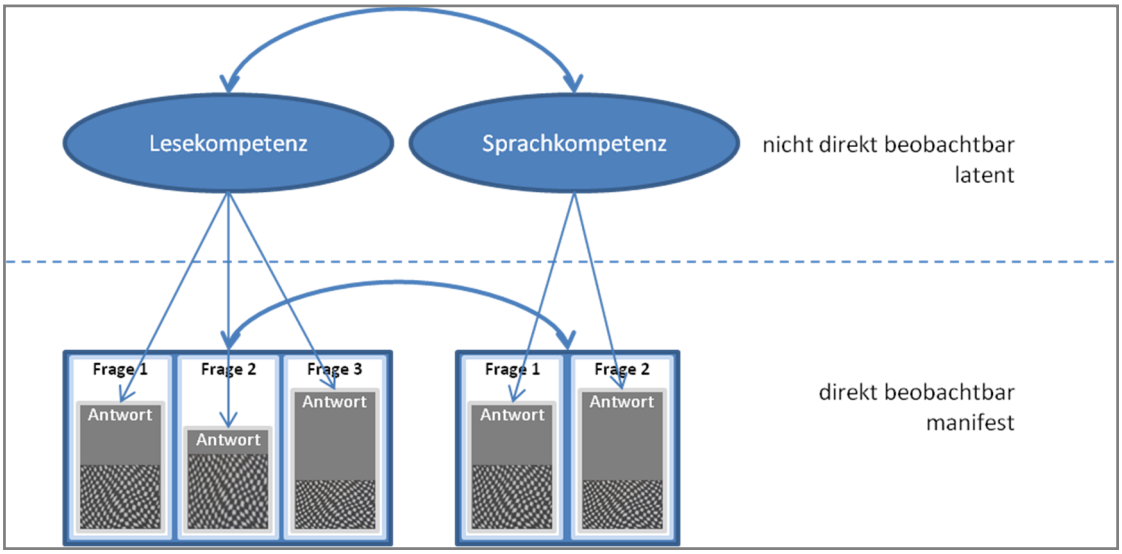


Abbildung 10: Konstruktvalidität

Beispiel: Konstruktvalidität

Aus der wissenschaftlichen Literatur ist bekannt, dass die Lesekompetenz von Schülerinnen und Schülern in engem Zusammenhang steht mit ihrer jeweiligen Sprachkompetenz. Die Annahme über den Zusammenhang zwischen den beiden latenten Personenmerkmalen „Lesekompetenz" und „Sprachkompetenz" kann im Rahmen einer konvergenten Validierung des Lesekompetenztests genutzt werden. Dafür setzen wir den Lesekompetenztest sowie einen Test zur Erfassung der Sprachkompetenz bei einer Stichprobe von Personen ein und berechnen hinterher die Korrelation zwischen den beiden Tests. Dieser empirisch beobachtete und durch den Korrelationskoeffizienten beschriebene Zusammenhang zwischen den beiden Tests sollte dem theoretisch angenommenen Zusammenhang zwischen den latenten Personenmerkmalen entsprechen. In unserem Beispiel würden wir also einen Korrelationskoeffizienten mit einem möglichst hohen positiven Betrag erwarten.

1.3.4.3 Objektivität

Die Objektivität eines Tests betrifft die Frage, in welchem Ausmaß ein Testergebnis abhängig ist vom Testanwender, also von der Person, die den Test durchführt (nicht zu verwechseln mit der Person, die getestet wird!). Es ist einleuchtend, dass die Antworten, die eine getestete Person in einem Test liefert, möglichst ausschließlich von dem betreffenden Merkmal der getesteten Person abhängen sollten, vollkommen unabhängig davon, wer mit ihr diesen Test durchführt. Bspw. sollten wir mit unserem Lesekompetenztest, den wir in einer Klasse einsetzen, bei jedem Schüler zu demselben Ergebnis gelangen, unabhängig davon, wer den Test in der Klasse durchführt, unabhängig davon, wer die Punkte für die niedergeschriebenen Antworten vergibt und unabhängig davon, wer aufgrund der Punktzahl entscheidet, ob einem Schüler eine hohe oder eine niedrige Lesekompetenz zugesprochen werden kann. Erreicht wird diese Unabhängigkeit vom Testanwender durch eine sogenannte *Standardisierung* des Tests. Sie ist gegeben, wenn in einer Testanleitung genau beschrieben wird, wie und unter welchen Bedingungen der Test durchgeführt werden muss, nach welchen Kriterien die Antworten im Test ausgewertet und mit Punkten versehen werden und wie die im gesamten Test erreichten Punkte (das Testergebnis) zu interpretieren sind (vgl. zum letzten Punkt Kap. 1.2). Entsprechend unterscheidet man auch zwischen der *Durchführungsobjektivität*, der *Auswertungsobjektivität* sowie der *Interpretationsobjektivität*.

Standardisierung

Durchführungsobjektivität

Die Durchführungsobjektivität ist gegeben, wenn ein Test immer auf dieselbe Art und Weise und unter denselben Bedingungen durchgeführt wird. Die Durchführungsobjektivität unseres Lesekompetenztests wäre bspw. gefährdet, wenn wir den Test einmal morgens während der ersten Schulstunde durchführten, ein anderes Mal jedoch in der letzten Nachmittagsstunde. Es wäre zu erwarten, dass die Testleistungen der Schüler in der ersten Stunde deutlich besser ausfallen als bei den Schülern, die am Ende eines langen Schultages erschöpft sind und entsprechend geringe Testleistungen zeigen. Die Bedingungen der Testdurchführung wären nicht miteinander vergleichbar und das Kriterium der Durchführungsobjektivität damit verletzt.

Auswertungsobjektivität

Die Auswertung der Testantworten, sprich die Entscheidung, wie viele Punkte der Schüler für eine Testantwort erhält, muss ebenfalls unabhängig sein von der Person, die diese Auswertung durchführt. Dabei gilt, je eindeutiger eine Antwort als richtig oder falsch, als gut oder schlecht bewertet werden kann, desto höher die Auswertungsobjektivität. Es gibt zwei Wege, diese Eindeutigkeit herzustellen. Zum einen können Aufgaben mit einem sogenannten „geschlossenen" Antwortformat eingesetzt werden. Beispiele hierfür wären die bekannten Multiple-Choice-Aufgaben oder Lückentexte. Geschlossene Antwortformate zeichnen sich dadurch aus, dass vorab genau definiert werden kann, wie eine gute bzw. richtige Antwort aussieht. Im Falle von Multiple-Choice-Aufgaben ist genau definiert, welche der präsentierten Optionen angekreuzt werden muss (und welche nicht), um die volle Punktzahl zu erreichen. Bei Lückentexten ist vorab für jede Lücke genau ein Wort definiert, welches in die Lücke zu schreiben ist, um den entsprechenden Punkt zu erhalten.

Geschlossene Aufgaben

Zum anderen können Auswertungsanleitungen und Schablonen erstellt werden, an die die auswertende Person sich möglichst genau hält. Genaue Auswertungsanleitungen sind insbesondere dann notwendig, wenn Aufgaben mit einem offenen Antwortformat eingesetzt werden. Offene Antwortformate zeichnen sich dadurch aus, dass es nicht genau eine richtige Antwort, sondern theoretisch unendlich viele richtige Antworten geben kann. Beispiel hierfür wären Fragen, die in Form eines mehr oder weniger langen Aufsatzes zu beantworten sind. In diesem Fall muss bei der Auswertung der Aufsatz nach fest vorgegebenen Kriterien (bspw. Argumentationsstruktur) und auf fest vorgegebene Art und Weise bewertet werden. Hierfür bieten sich z.B. Checklisten an, in denen die verschiedenen Kriterien als erfüllt oder nicht erfüllt abgehakt werden können.

Offene Aufgaben

Interpretationsobjektivität

Die Interpretationsobjektivität kann erhöht werden, indem Interpretationshilfen und -anweisungen zur Verfügung gestellt werden. Solche Hilfestellungen sind bspw. Normtabellen, wie wir sie in Kap. 1.2.2 unter dem Stichwort „Soziale Bezugsnorm" besprochen haben. Sie stellen einen Vergleichsmaßstab zur Verfügung, anhand dessen das Testergebnis einer Person in Bezug auf eine vergleichbare Gruppe von Personen bewertet und interpretiert werden kann. Interpretationshilfen in Bezug auf eine kriteriale Bezugsnorm könnten bspw. inhaltliche Beschreibungen verschiedener Abschnitte der Testskala sein. Auf diese Form wird im Studienbrief „Vergleichsarbeiten" in Kapitel 2.4 detailliert eingegangen. Derartige Interpretationshilfen standardisieren die Testinterpretation und versuchen so, zu verhindern, dass individuelle Deutungen die Interpretation des Testergebnisses beeinflussen.

1.3.5 Nebengütekriterien

Die drei klassischen Testgütekriterien der Reliabilität, Validität und Objektivität werden durch sogenannte Nebengütekriterien ergänzt. Nebengütekriterien zeichnen sich durch einen starken Anwendungsbezug aus. Die Nebengütekriterien sind im Gegensatz zu den klassischen Gütekriterien uneinheitlich definiert. Die folgenden Kriterien werden in der Literatur häufig unter diesem Begriff zusammengefasst: Skalierung, Normierung, Testökonomie, Nützlichkeit, Unverfälschbarkeit und Fairness. Eine detaillierte Beschreibung dieser Kriterien findet sich bei Moosbrugger und Kevala (2008), die im Übrigen auf die begriffliche Abgrenzung zwischen Haupt- und Nebengütekriterien verzichten. Obwohl sämtliche der aufgeführten Nebengütekriterien relevant sind, würde deren Besprechung über den Rahmen dieses Kapitels hinausgehen. Daher soll in den folgenden Abschnitten einzig auf das Kriterium der Normierung eingegangen werden, welches für die Individualdiagnostik von zentraler Bedeutung ist.

1.3.5.1 Normierung

Die Ergebnisse eines Tests werden zunächst als Rohwert (Score) angegeben. Im Falle eines Leistungstests wäre dies bspw. die Anzahl der korrekt beantworteten Aufgaben. Um die Bedeutung dieser Rohwerte zu ermitteln, ist es erforderlich, die Ergebnisse in Relation zu den Ergebnissen einer vergleichbaren Referenzgruppe zu setzen (vgl. Kap. 1.2.2). Die Normierung eines Tests gibt auf Grundlage einer umfangreichen, repräsentativen Normierungsstichprobe Auskunft über die übliche Ausprägung des untersuchten Merkmals innerhalb einer Referenzgruppe. Die Testnormen können üblicherweise dem Testmanual bzw. den Handreichungen entnommen werden. Sollten bestimmte Faktoren wie Geschlecht und Alter für die Merkmalsausprägung relevant sein, so kann die Referenzgruppe anhand dieser Kriterien ausdifferenziert werden. In diesem Fall stehen entsprechende Geschlechts-, Alters- oder Schulnormen zur Verfügung. Die Normierungsdaten eines Tests bieten somit einen Bezugsrahmen zur Interpretation

von Testergebnissen und ermöglichen durch den Vergleich mit einer Referenzgruppe eine Standortbestimmung einzelner Schüler.

Definition: Normierung

Das Nebengütekriterium der Normierung bewertet, inwieweit für die Ergebnisse eines Testinstruments Vergleichsdaten vorhanden sind, anhand derer sich Einzelergebnisse interpretieren lassen (Rost, 2004). Rost betont, dass ein normiertes Testinstrument nicht zwangsläufig auch den Hauptgütekriterien entspricht. Das Gütekriterium der Normierung ist unabhängig von der Objektivität, Reliabilität und Validität eines Instruments.

Normtabellen enthalten meist auch Angaben darüber, welchem Prozentrangplatz ein Testergebnis entspricht. Anhand des Prozentrangplatzes kann man ablesen, wie viele der Personen in der Normierungsstichprobe ein besseres und wie viele ein schlechteres Testergebnis erzielt haben. Bspw. bedeutet ein Prozentrangplatz 87, dass 13% der Personen in der Normierungsstichprobe ein höheres Testergebnis erzielt haben. Bei einem Prozentrangplatz von 50 hätten genauso viele Personen ein höheres Ergebnis wie Personen ein niedrigeres Ergebnis, bei einem Prozentrangplatz von 20 zeigten 80% der Normierungsstichprobe eine höhere Testleistung.

1.3.6 Weiterführende Literatur

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2. Aufl.). München: Pearson Studium.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6. Aufl.). Weinheim: Beltz.

Krauth, J. (1995). *Testkonstruktion und Testtheorie*. Weinheim: Beltz.

Moosbrugger, H. & Kelava, A. (Hrsg.). (2008). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer.

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

1.3.7 Verständnis- und Diskussionspunkte

1. *Diskutieren Sie mögliche Ursachen, aus denen der Messfehler resultieren kann.*
2. *Unter welchen (ggf. auch unrealistischen) Bedingungen wäre die Reliabilität eines Tests gleich 1?*
3. *Was bedeutet „erwartungskonforme Korrelation" im Hinblick auf die Validität eines Tests?*
4. *Welche Vorteile bietet die Odds-Even-Reliabilität (gerade-ungerade-Reliabilität) gegenüber der Split-Half- Reliabilität?*
5. *Diskutieren Sie, warum die Normierung eines Tests ein wichtiges Nebengütekriterium darstellt.*

1.4 Inhaltlicher Anwendungsbereich/Phänomenbereich

In diesem Kapitel soll eine Auswahl an Personenmerkmalen vorgestellt werden, die sich mit Hilfe standardisierter Testverfahren erfassen lassen. Dabei nehmen wir zwei Beschränkungen vor. Zum einen beschränken wir uns auf Personenmerkmale, die für pädagogische und schulische Kontexte relevant sind. Alle Personenmerkmale, die eher im Rahmen klinischer Fragestellungen interessant sind oder aber keinen Beitrag für primär pädagogische Entscheidungen leisten, werden in diesem Kapitel nicht behandelt. Außerdem teilen wir die verschiedenen Personenmerkmale in einige wenige Klassen ein und besprechen diese Klassen dann anhand eines prototypischen Beispiels. Auf diese Weise wird die enorme Menge an Personenmerkmalen auf ein überschaubares Maß gebracht. Eine weitere Einschränkung erfolgt aus didaktischen Gründen: Wir beschreiben ausschließlich Personenmerkmale, die sich durch etablierte und standardisierte Testverfahren diagnostizieren lassen. Der Sinn dahinter ist, dass in diesem Zuge „best practice"-Beispiele präsentiert werden, die die testtheoretischen Inhalte des letzten Kapitels 1.3 auf eine gute Weise wiederholen und illustrieren. Neben diesem Wiederholungseffekt soll es in diesem Kapitel jedoch vornehmlich um folgende Fragen gehen:

- *Welche Schülermerkmale sind von besonderer Bedeutung für schulische Leistung?*
- *Wie können Schulleistung und schulleistungsrelevante Merkmale beurteilt werden?*

In diesem Kapitel wenden wir uns zwei Klassen von Personenmerkmalen zu, die im schulisch-pädagogischen Kontext von besonderer Bedeutung sind. In Anlehnung an Hosenfeld und Schrader (2006) unterscheiden wir zwischen Schulleistungsmerkmalen und (schul-)leistungsrelevanten Merkmalen (Abbildung 11). Schulleistungsmerkmale lassen sich durch Schulleistungstests erfassen, die die Leistungsfähigkeit eines Schülers in einem oder mehreren Schulfächern testen. Schulleistungsrelevante Merkmale hingegen umfassen sowohl kognitive als auch motivationale Personeneigenschaften, die nicht direkt als ein Aspekt von Schulleistungsfähigkeit angesehen werden können, die jedoch eine Voraussetzung für Schulleistung sind. Bspw. ist eine angemessene Motivation eine notwendige Voraussetzung für jegliche Art von Leistung. Eine präzise und eindeutige Zuordnung einzelner schulleistungsrelevanter Merkmale in die Kategorien „kognitiv" oder „motivational" ist nicht immer möglich (und auch nicht notwendig). Daher sollte die Einteilung auch nicht als Dichotomie verstanden werden. Vielmehr bezeichnen wir damit die Pole einer gemeinsamen Dimension.

Eine Besprechung sämtlicher schulleistungsrelevanter Merkmale würde den Rahmen des Studienbriefs sprengen, daher behandeln wir, wie oben bereits angekündigt, exemplarisch die kognitiven und motivationalen Merkmale, die zum einen Bedingungen für die Schulleistung darstellen und zum anderen einen hohen Praxisbezug aufweisen. Wie aus Abbildung 11 hervorgeht, sind dem kognitiven Pol schulleistungsrelevanter Merkmale die Intelligenz und das Vorwissen zugeordnet. Da der Bereich der Intelligenz sehr umfassend ist, wird in diesem Kapitel ausschließlich dieser als Vertreter kognitiver Merkmale besprochen. Die Erfassung von Wissen und Vorwissen ist dann Gegenstand des Kapitels 1.5. Die Merkmale Selbstkonzept und Selbstwirksamkeitserwartung beinhalten sowohl kognitive als auch motivationale Aspekte und sind aus diesem Grund in der Mitte der Dimension angesiedelt. Im Studienbrief wird jedoch zunächst auf das Merkmal Selbstkonzept eingegangen. Anschließend wird die Diagnose der Motivation besprochen.

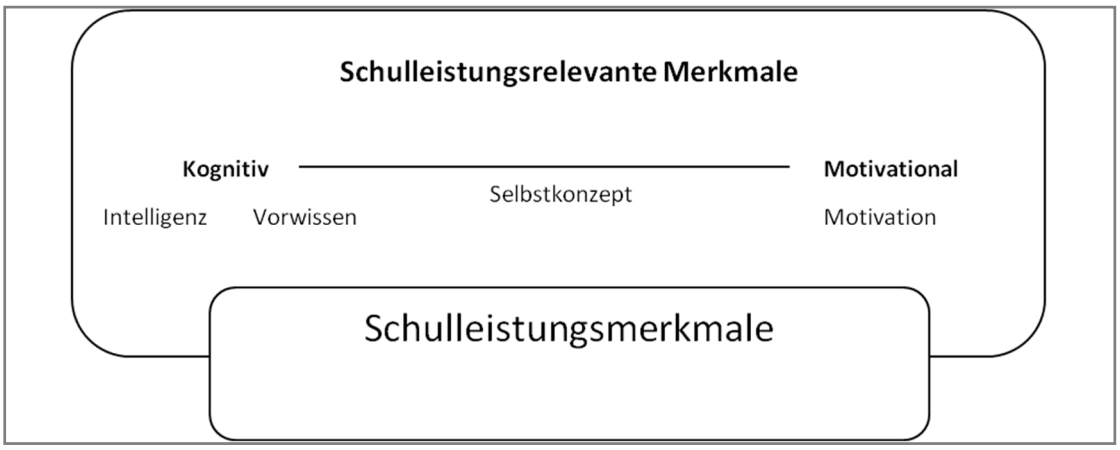


Abbildung 11: Schulisch-pädagogisch relevante Personenmerkmale

1.4.1 Schulleistungsmerkmale

Schulleistung wird im schulischen Alltag durch Zensuren quittiert. Heller (1984) definiert Schulleistung als „das gesamte Leistungsverhalten im Kontext schulischer Bildungsbemühungen." Rindermann und Kwiatkowski (2010) beschreiben Schulleistung als die „individuelle Leistung eines Schülers und daraus abgeleitet seine kognitiven Schulfähigkeiten (Wissen und Verständnis oder Denken und Wissen), indirekt auch Intelligenz und andere schulleistungsrelevante Personen- und Umweltmerkmale bis zu Erziehungsstilen und Bildungsorientierung der Eltern." Diese Definition verdeutlicht, dass Schulleistung sehr viele Facetten hat und von vielen Faktoren beeinflusst wird. Einer dieser Faktoren sind sicherlich die Kenntnisse und Fähigkeiten eines Schülers. Schulleistung ist allerdings mit Kenntnissen und Fähigkeit nicht gleichzusetzen: Die erforderliche Fähigkeit kann zwar vorhanden sein, aber im entscheidenden Moment, bspw. einer Prüfungssituation, nicht abgerufen werden. Daher unterscheidet man zwischen Performanz und Potenzial. Kenntnisse und Fähigkeiten stellen ein Potenzial dar, während Performanz als „Ausdruck oder die Anwendung von Fähigkeiten in lebensweltlich relevanten Situationen, etwa in der Schule oder allgemeiner in Ausbildungssituationen, beim Lernen, im Wissenserwerb und in Prüfungen, in der beruflichen Tätigkeit oder in Anforderungssituationen des Alltagslebens" definiert wird (Rindermann & Kwiatkowski, 2010). Schulleistung kann im Sinne einer Performanz abgegrenzt werden von Personenmerkmalen wie z.B. Intelligenz o.ä., die im Sinne eines Potenzials zu interpretieren sind. Um diese Abgrenzung sprachlich zu untermauern, soll im vorliegenden Studienbrief zwischen „Perfor-

Potenzial vs. Performanz

manz“ (Schulleistung) und „Potenzial“ (kognitive Grundfähigkeiten) unterschieden werden. Zur Erfassung schulischer Performanz stehen standardisierte Diagnoseinstrumente zur Verfügung, die als Schulleistungstests bezeichnet werden.

1.4.1.1 Schulleistungstests

Schulleistungstests lassen sich im Hinblick auf Messintention und Durchführungsbedingungen unterscheiden. Die Messintention eines Leistungstests kann sich auf die Erfassung allgemeiner Schulleistung oder aber auf eine bestimmte Teilleistung richten. Zudem wird zwischen bezugsgruppenorientierten und kriteriumsorientierten Schulleistungstests differenziert (vgl. Kap. 1.2). Bezugsgruppenorientierte Schulleistungstests sind Tests, bei denen das individuelle Ergebnis mit den an einer relevanten Stichprobe (meist Klassenstufe) ermittelten Ergebnissen verglichen wird. Ein kriteriumsorientierter Test dagegen ist ein wissenschaftliches Routineverfahren zur Untersuchung der Frage, ob und eventuell wie gut ein bestimmtes Lehrziel erreicht ist. Die hierbei verwendeten Testaufgaben sind nicht identisch mit dem Lehrziel, sondern repräsentieren es nur und dienen dazu, den individuellen Fähigkeitsgrad eines Schülers mit einem gewünschten Fähigkeitsgrad zu vergleichen“ (Fricke, 1973; Ingenkamp & Lissmann, 2008).

Definition: Schulleistungstests

Schulleistungstests sind Verfahren der Pädagogischen Diagnostik, mit deren Hilfe Ergebnisse geplanter und an Curricula orientierter Lernvorgänge möglichst objektiv, zuverlässig und gültig gemessen und durch Lehrende oder Beratende ausgewertet, interpretiert und für pädagogisches Handeln nutzbar gemacht werden können (Ingenkamp & Lissmann, 2008).

Unterschieden werden kann weiterhin zwischen Mehrfächertests, welche die Schulleistung in relevanten Fächern überprüfen und Tests, die auf die Erfassung der Leistung in einem bestimmten Bereich (bspw. Mathematik oder Leseverständnis) abzielen. Letztere eignen sich insbesondere zur Diagnose von Teilleistungsstörungen und werden daher im entsprechenden Abschnitt behandelt. Mehrfächertests finden ihre Anwendung in der Schullaufbahnberatung, wo sie eine datengestützte Prognose für die schulische Leistungsentwicklung ermöglichen. Darüber hinaus können Schulleistungstests zur Überprüfung des Vorwissens eingesetzt werden. Im vorliegenden Abschnitt soll stellvertretend für die Mehrfächertests der Hamburger Schulleistungstest für 4. und 5. Klassen vorgestellt werden.

1.4.1.2 Hamburger Schulleistungstest für vierte und fünfte Klassen

Diagnoseziel

Der Hamburger Schulleistungstests (HST) gehört zu den sogenannten „Mehrfächertests“ und erfasst verschiedene relevante Aspekte schulischen Lernens. Um die angestrebte Lehrzielvalidität zu gewährleisten, wurde der HST laut Testbeschreibung unter Berücksichtigung der curricularen Anforderungen der Klassen 4 und 5 konstruiert. Der Test umfasst 14 Untertests, die auf 5 Inhaltsbereiche (Subskalen) verteilt sind. Besonderes Merkmal des HST ist, dass die Informationsentnahme aus Karten, Tabellen und Diagrammen als Indikator für Schulleistung verstanden wird. Im Hinblick auf die zunehmende Relevanz selektiver Informationsentnahme, bspw. aus dem Internet, scheint die Erfassung dieser Kompetenz als sinnvoll. Da der HST sehr umfangreich ist, sollte er an zwei Tagen (eine Doppelstunde am ersten Tag und eine Einzelstunde am Folgetag) durchgeführt werden.

Aufgabe: Datenbankrecherche

Besuchen Sie im Internet die UDiKom-Testdatenbank (http://tests.udikom.de/). Suchen Sie dort den Hamburger Schulleistungstest. Welche fünf Inhaltsbereiche werden durch den HST abgedeckt?

1. Subskala: _____
2. Subskala: _____
3. Subskala: _____
4. Subskala: _____
5. Subskala: _____

Validität des HST

Hinweise auf Kriteriumsvalidität ergeben sich u.a. aus einer signifikanten Korrelation mit dem Notendurchschnitt (r = -0,73). Betrachtet man die Korrelation zwischen der Mathematiknote und den Ergebnissen des HST im Bereich Mathematik, ist eine etwas niedrige Validität festzustellen (r = -0,57). Die

negative Korrelation ergibt sich aus dem Umstand, dass das Verhältnis zwischen Schulnote und Leistung nicht dem Verhältnis zwischen Testergebnis und Leistung entspricht: während eine „hohe“ (schlechte) Note wie 5 einer mangelhaften Leistung entspricht, beschreibt ein „hohes“ Testergebnis eine gute Leistung. So entsteht eine erwartungskonforme, aber negative Korrelation: Je besser (höher) der Testscore, desto besser (niedriger) die Note.

Korrelation mit Noten

1.4.1.3 Was ist ein „guter” Schulleistungstest?

Um die Zweckmäßigkeit und die Qualität eines Schulleistungstests zu bewerten, schlägt Langfeldt (1984) folgende „Prüfsteine“ vor. Die „Prüfsteine“ beschränken sich dabei keinesfalls ausschließlich auf Schulleistungstests, sondern können auch als Bewertungskriterien für andere Testverfahren herangezogen werden.

Qualitätskriterien

1. Überprüft der Test das, was unterrichtet wurde?
2. Ist der Test reliabel (zuverlässig) genug?
3. Wie präzise ist ein individueller Testpunktwert?
4. Wie wird eine objektive Testdurchführung gesichert?
5. Wie wird die Auswertungsobjektivität gewährleistet?
6. Wie ist der Test normiert?
7. Gibt es Paralleltests?
8. Wie sind die Ergebnisse inhaltlich zu interpretieren?
9. Wie lange dauert der Test?
10. Wie alt ist der Test?

Darüber hinaus bietet Langfeldt (1984) eine Orientierungshilfe zur Einschätzung der Testqualität auf Basis der angegebenen Testkennwerte. So sollten „brauchbare“ Schulleistungstests einen Validitätswert von mindestens $r = 0{,}60$ (wenn ein Zusammenhang und nicht eine Unabhängigkeit vermutet wurde, vgl. Kap. 1.3.4.2) und einen Reliabilitätskoeffizienten von mindestens Cronbachs $\alpha = 0{,}80$ aufweisen. Die Normierungsstichprobe sollte mindestens 500 Personen umfassen.

1.4.2 Schulleistungsrelevante Merkmale: Intelligenz

Intelligenz ist sicherlich eines der bedeutsamsten Personenmerkmale, die Schulleistung beeinflussen. Ihre Diagnostik erfordert allerdings vertiefte methodische Kenntnisse und sollte daher von entsprechend ausgebildeten Psychologen durchgeführt werden. Entsprechend dienen die folgenden Abschnitte nicht der Befähigung zur Durchführung von Intelligenztests. Das angestrebte Ziel ist vielmehr die Herstellung von Transparenz, um Intelligenzdiagnostik, die bspw. im Rahmen der Hochbegabtendiagnostik durchgeführt wird, und ihre Ergebnisse nachvollziehen und bewerten zu können.

Was Intelligenz genau ist, ist eine Frage, die in der Literatur unterschiedlich beantwortet wird und zu einer großen Anzahl unterschiedlicher Definitionen geführt hat. An dieser Stelle soll diese Diskussion um die „richtige“ Definition nicht abgebildet werden. Stattdessen wählen wir einfach eine Definition, wie sie von Amelang und Schmidt-Atzert (2006) angeboten wird und eine gleichermaßen prägnante und praxisrelevante Definition von Intelligenz darstellt.

Intelligenz – Was ist das?

Definition: Intelligenz

Unter Intelligenz wird das Potenzial einer Person verstanden, kognitive Leistungen zu erbringen. Eine hoch intelligente Person kann, muss aber nicht gute Leistungen in Schule und Beruf zeigen. Motivationale Gründe oder ungünstige Arbeitsbedingungen können dazu führen, dass die Person nicht die Leistung erbringt, zu der sie eigentlich fähig wäre.

Das Brickenkamp-Testkompendium allein umfasst 57 verschiedene Testinstrumente zur Intelligenzmessung. Auf der Homepage des Testverlags Hogrefe sind 27 Intelligenztests für Kinder und Jugendliche aufgeführt. Die Vielzahl der verfügbaren Testinstrumente ist ein Indikator für die Popularität von Intelligenztests, welche sich z.B. durch die Fähigkeit von Intelligenztests erklären lässt, zuverlässig Niedrig- und Hochbegabung zu diagnostizieren. In der Tat gelten Intelligenztests daher als die wohl erfolgreichsten psychologischen Diagnoseverfahren (Amelang & Schmidt-Atzert, 2006). Ein wesentliches Merkmal

zur Unterscheidbarkeit von Intelligenztests ist die Messintention, also die Frage, wie Intelligenz theoretisch durch ein Modell beschrieben wird (vgl. 1.3.4.2). Während einige Testinstrumente wie bspw. das Leistungsprüfsystem (LPS) ein Intelligenzmodell, das viele verschiedene Dimensionen beschreibt, durch sehr heterogene Untertests abbilden, erfassen andere Testinstrumente Intelligenz mit nur einer Skala als allgemeines intellektuelles Gesamtpotenzial, welches auch als Grundintelligenz oder g-Faktor der Intelligenz bezeichnet wird. Andere Verfahren sind an der Ausprägung spezifischer Intelligenzfaktoren interessiert.

Klassifikation von Intelligenztests

Im vorliegenden Studienbrief werden Intelligenztests anhand dreier Merkmale klassifiziert. Das Erste betrifft die Frage nach der Anzahl der *Dimensionen*, die das zu Grunde gelegte theoretische Modell der Intelligenz beschreibt. Das zweite Merkmal betrifft die *Sprachfreiheit* der Tests. Schließlich wird danach unterschieden, ob der Test unter strikten *Zeitvorgaben* durchgeführt werden muss oder nicht.

Dimensionalität

Bezüglich der Dimensionalität unterscheiden sich Testinstrumente dahingehend, ob der Test die allgemeine Intelligenz (general factor) „g" oder einen oder mehrere Aspekte von Intelligenz erfasst. Es geht also darum, ob Intelligenz eindimensional als ein übergreifendes, globales Konstrukt verstanden wird oder ob verschiedene „Intelligenzen" angenommen werden. Eindimensionale Intelligenztests sind rasch durchführbar und liefern eine globale Einschätzung des intellektuellen Potenzials. Mehrdimensionale Tests dagegen ermöglichen die Erfassung eines kognitiven Profils mit speziellen Stärken und Schwächen.

Sprachgebundenheit

Die Frage nach der Sprachfreiheit stellt sich insbesondere, wenn Schüler mit einer anderen Muttersprache getestet werden sollen. Soll bspw. die Intelligenz eines Schülers mit Migrationshintergrund überprüft werden, der die jeweilige Landessprache noch nicht beherrscht, so eignet sich ein sprachfreier Test um sprachbedingte Benachteiligung auszuschließen.

Zeitvorgaben

Intelligenztests unterscheiden sich nicht nur in Bezug auf die theoretischen Merkmale, sondern auch nach den Durchführungsbedingungen, bspw. ob der Test in Einzeltestung oder in der Gruppe erfolgt. Ein Gruppentest kann aus ökonomischen Gründen sinnvoll sein. Aus motivationalen Gründen ist es jedoch teilweise ratsam, eine Einzeltestung durchzuführen. Das ist insbesondere dann der Fall, wenn es sich um Testpersonen mit kognitiven Beeinträchtigungen oder um die Diagnose besonderen Förderbedarfs handelt. Intelligenztests lassen sich zudem in sogenannte „Speedtest" und „Powertests" aufteilen. Bei Speedtests sind enge zeitliche Vorgaben für die Testbearbeitung gegeben, die zu einer Belastung während der Testbearbeitung führen (sollen). Bei Powertests entfällt dieser Zeitdruck. Die Entscheidung, ob der Test eine starke zeitliche Begrenzung vorgeben sollte oder nicht, sollte davon abhängig gemacht werden, ob die Zeitbegrenzung in Kombination mit Faktoren wie Leistungsängstlichkeit oder Sprachproblemen des Schülers zu verzerrten Ergebnissen führen kann. Trifft dies zu, sollte die Entscheidung zugunsten Powertests ausfallen, bei denen zwar keine eng bemessene Zeitbegrenzung vorgegeben wird, bei denen jedoch die Schwierigkeit der Items graduell ansteigt.

In den folgenden Abschnitten sollen einige ausgewählte Intelligenztests vorgestellt und besprochen werden.

1.4.2.1 Eindimensionale Intelligenztests: Der Hamburg-Wechsler-Intelligenz-Test

Diagnoseziel

Um die Kategorie eindimensionaler Intelligenztests zu veranschaulichen, soll an dieser Stelle der Hamburg-Wechsler Intelligenztest (HAWIK) vorgestellt werden (für weiterführende Informationen siehe http://tests.udikom.de/). Dieser enthält zwar 13 verschiedene Untertests (Tabelle 2), weshalb man annehmen könnte, dass er ein mehrdimensionales Modell von Intelligenz abbilden soll. Aus Wechslers Definition der Intelligenz wird jedoch deutlich, dass er nicht verschiedene „Intelligenzen" unterscheidet, sondern Intelligenz als ein globales Potenzial versteht.

Definition: Intelligenz

Wechsler (1964) definiert Intelligenz als „globale oder zusammengesetzte Fähigkeit des Individuums, zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umgebung wirkungsvoll auseinanderzusetzen."

Der Test spiegelt das Intelligenzmodell in Form von 13 Untertests (Subtests) wider, die sich in einen Handlungsteil und einen Verbalteil gliedern und in einer festgelegten Reihenfolge durchgeführt werden. Daraus folgt, dass Intelligenz gemäß des HAWIK als die Summe verbaler und praktischer Fähigkeiten operationalisiert wird, die durch die Untertests gemessen werden. Diese Summe ist jedoch zunächst wenig aussagekräftig (vgl. Kap.1.2.2), sondern ist in Bezug auf eine entsprechende Altersnorm zu interpretieren. Der Test liefert differenzierte Altersnormen. Dies ist von besonderer Bedeutung für einen auf Kinder und Jugendliche ausgerichteten Intelligenztest. Der HAWIK–III wird diesem Anspruch gerecht,

da er in Altersgruppen gestaffelt ist, die sich jeweils um nur 4 Monate unterscheiden. Die Größe der Stichprobe innerhalb einer Alterskohorte liegt dadurch allerdings nur zwischen 35 und 60 Personen.

Abkürzung	Untertest	Beispielaufgabe
AW	Allgemeines Wissen	In welcher Himmelsrichtung geht die Sonne unter?
GF	Gemeinsamkeiten finden	Was ist das Gemeinsame an Hemd und Schuh?
RD	Rechnerisches Denken	Franz liest 3 Seiten in 5 Minuten. Wie viele Minuten braucht er für 24 Seiten?
WT	Wortschatz-Test	Was ist ein Brot?
AV	Allgemeines Verständnis	Warum haben Autos Sicherheitsgurte?
ZN	Zahlen nachsprechen	3-4-1-7
BE	Bilder ergänzen	Was fehlt auf dem Bild? Fehlende Details benennen oder zeigen
ZS	Zahlen-Symbol-Test	Umwandlungstabelle mit Zahlen und Symbolen (z.B. +). Symbole in Felder und Zahlen eintragen.
BO	Bilder ordnen	Bilder in die richtige Reihenfolge bringen
MO	Mosaik-Test	Zweifarbiges Muster mit 2, 4 bzw. 8 Klötzchen nachlegen
FL	Figurenlegen	Zerschnittene Figuren („Puzzle“) zusammenfügen
SS	Symbolsuche	Zwei Gruppen von Symbolen vorgegeben. Ankreuzen, ob ein Symbol in beiden Gruppen enthalten ist
LA	Labyrinthtest	Linie vom Zentrum zum Ausgang eines Labyrinths ziehen

Tabelle 2: Untertests des HAWIK-III

Der Vergleich des Testergebnisses mit der jeweiligen Altersnorm führt – wie bei anderen Intelligenztests auch – zur Berechnung des sogenannten „Intelligenzquotienten“.

Definition: Intelligenzquotient

Der Intelligenzquotient (IQ) zeigt das intellektuelle Leistungsvermögen einer Person im Vergleich zu einer Normstichprobe vergleichbaren Alters an. Üblicherweise wird das durchschnittliche intellektuelle Leistungsniveau einer Altersgruppe auf 100 IQ-Punkte festgelegt. Eine Standardabweichung beträgt üblicherweise 15 IQ-Punkte.

Durchführung
Auswertung
Interpretation

Durch die zahlreichen, recht komplexen Untertests, fordert die Durchführung des HAWIK–III viel Übung seitens des Testleiters. Die Aufgaben werden mit Hilfe einer Lösungsschablone ausgewertet, und die Rohwerte werden addiert. Die Summe ergibt die Punktezahl für den jeweiligen Untertest. Ein spezielles Programm übernimmt die Auswertung der Rohwerte, einschließlich der Ermittlung des Intelligenzquotienten und einer graphischen Darstellung der individuellen Intelligenzprofile. Zur Interpretation der Ergebnisse, insbesondere zur Erklärung schwacher Subtestergebnisse, bietet der HAWIK differenzierte Informationen. Amelang und Schmidt-Atzert (2006) sehen dies als zentralen Vorteil des Instrumentariums: „Der HAWIK–III stellt trotz einiger kleiner Unzulänglichkeiten ein brauchbares und nützliches Intelligenztestverfahren für Kinder und Jugendliche dar. Die Informationsausbeute ist groß. Der Test liefert neben dem Intelligenzquotienten viele Informationen über die Stärken und Schwächen des Probanden“.

Aufgabe: Datenbankrecherche

Der HAWIK-III liefert einen IQ-Wert, der das intellektuelle Potenzial von Kindern oder Jugendlichen ausdrücken soll. Er kann jedoch auch genutzt werden, um Stärken und Schwächen in vier verschiedenen Teilleistungsbereichen zu identifizieren. Um welche Teilleistungsbereiche handelt es sich (siehe http://tests.udikom.de/)?

1.4.2.2 Mehrdimensionale Intelligenztests: Das Leistungsprüfsystem

Diagnoseziel

Das Leistungsprüfsystem (LPS) spiegelt das Intelligenzmodells von Thurstone (1938/1947) wider. Dieses Modell beschreibt Intelligenz mehrdimensional, indem es 7 unabhängige sogenannte „Primärfaktoren" der Intelligenz postuliert. Jeder Primärfaktor wird mit mindestens zwei Untertests bzw. 80 Aufgaben getestet. Das LPS eignet sich insbesondere dann, wenn eine möglichst differenzierte Aussage über verschiedene kognitive Fähigkeiten vorliegen soll.

Aufgabe: Datenbankrecherche

Suchen Sie unter http://tests.udikom.de/ das Leistungsprüfsystem und sammeln Sie dort Informationen über die verschiedenen Untertests. Wie würden Sie die sieben Primärfaktoren benennen, die durch diese 14 Untertests abgebildet werden sollen?

1.4.2.3 Sprachfreie Tests: Der Culture Fair Test 20

Sprachfreie Tests werden vornehmlich dann eingesetzt, wenn (mutter-)sprachliche oder sozio-kulturelle Einflüsse auf die Testleistung vermieden werden sollen. Sie werden daher auch als „kulturfreie" Tests bezeichnet. Jedoch gibt es bislang noch keinen Intelligenztest, der die Intelligenz vollkommen unabhängig von sozio-kulturellen Einflüssen messen könnte. Daher wird mittlerweile eher von „kulturfairen" Tests gesprochen.

Diagnoseziel

Sprachfreie Tests basieren auf dem Intelligenzkonzept von Cattell (1968), welches zwischen fluider und kristalliner Intelligenz unterscheidet. Die kristalline Intelligenz bezieht sich auf die „Sammlung gelernter Kenntnisse, die sich ein Mensch angeeignet hat, in dem er seine fluide Intelligenz in der Schule anwandte" (Cattell & Piaggio, 1973). Diese kristallinen Intelligenzanteile sind entsprechend stark von sprachlichen Fähigkeiten und der Kultur abhängig. Fluide Intelligenz dagegen bezeichnet bildungsunabhängige, intellektuelle Fähigkeiten wie „die Fähigkeit komplexe Beziehungen in neuartigen Situationen wahrnehmen und erfassen zu können" (Cattell, 1968). Sprachfreie bzw. kulturfaire Tests, wie z.B. der Culture Fair Test 20 (CFT 20), zielen daher meist auf die Erfassung fluider Intelligenzanteile ab.

Durchführung Auswertung

Der CFT 20 ist auf die Erfassung der fluiden Intelligenz ausgerichtet und umfasst ausschließlich sprachfreie Aufgaben, die unabhängig von erlerntem Wissen und daher kulturfrei sind. Der CFT 20 setzt sich aus 4 Untertests mit insgesamt 92 Aufgaben zusammen, die es erfordern, Figurenreihen fortzusetzen, Figuren zu klassifizieren, Figurenmatrizen zu vervollständigen und topologische Schlüsse ziehen. Innerhalb eines Untertests sind die Aufgaben nach Schwierigkeit gestaffelt. Die Aufgaben sind auf zwei identische Testformen verteilt, wobei der erste Teil auch als Kurzversion eingesetzt werden kann. Der CFT 20 kann als Gruppentest aber auch in Einzeltestung mit Testpersonen im Alter von 8-70 Jahren durchgeführt werden. Jeder Untertest beginnt zunächst mit Einführungsaufgaben, um die Testpersonen mit den Anforderungen des jeweiligen Untertests vertraut zu machen. Die Durchführungszeit beträgt ca. 55 Minuten (Kurzversion 35 Minuten), was eine verhältnismäßig rasche und ökonomische Einschätzung der Grundintelligenz ermöglicht. Auch die Auswertung der Testergebnisse gestaltet sich durch eine verfügbare Auswertungsschablone als objektiv und zeitlich ökonomisch.

1.4.3 Schulleistungsrelevante Merkmale: Motivation

Theoretischer Hintergrund

Intelligenz von Schülern, wie wir sie eben besprochen haben, ist zwar eine notwendige, aber noch lange keine hinreichende Bedingung für schulischen Erfolg. Schulerfolg wird durch das Zusammenspiel verschiedener Faktoren beeinflusst. Als wichtiger Bedingungsfaktor gilt hierbei auch die Motivation. Motivation hat sich im Hinblick auf Lernen und Leistung fest im erziehungswissenschaftlichen Diskurs sowie in der Alltagssprache etabliert. Sie umfasst nach Langfeldt (2006) all diejenigen Prozesse, die zielgerichtete Verhaltensweisen in konkreten Situationen auslösen und aufrechterhalten. Der Motivation werden bestimmte lernleistungsförderliche Funktionen zugeschrieben. So geht bspw. Ormond (2006) davon aus, dass Motivation zur Verbesserung kognitiver Prozesse und zur Leistungssteigerung beiträgt. Weiterhin hat Motivation einen Einfluss darauf, was als zufriedenstellend empfunden wird und bestimmt daher auch Verhaltensabsichten und Intentionen. Bezieht sich die Motivation auf Schulleistung, so kann zwischen Lernmotivation und Leistungsmotivation unterschieden werden.

Definition: Lernmotivation

Lernmotivation bezeichnet die Form der Motivation, welche die Absicht oder die Bereitschaft einer Person beschreibt, sich in einer konkreten Situation mit einem Gegenstand lernend auseinander zu setzen (Wild et al., 2001).

Definition: Leistungsmotivation

Leistungsmotivation bezeichnet die Absicht, etwas zu leisten, Erfolge zu erzielen und Misserfolge zu vermeiden, wobei zur Bewertung des Erfolges bzw. Misserfolgs ein individuell verbindlicher Bewertungsmaßstab herangezogen wird (Wild et al., 2001). Auf Basis dieser Definition kann zwischen den Dimensionen „Annäherung an Erfolg" und „Vermeidung von Misserfolg" unterschieden werden (Langfeldt, 2006). Schüler sind motiviert durch die Absicht, Erfolge zu erreichen oder Misserfolg zu vermeiden.

1.4.3.1 Erfassung der Lern- und Leistungsmotivation: Der SELLMO

Diagnoseziel

Zur Diagnose der Lern-Leistungsmotivation kann bspw. auf die Skalen zur Erfassung der Lern- und Leistungsmotivation (SELLMO) von Dickhäuser et al. (2002) zurückgegriffen werden. Die Skalen basieren auf der theoretischen Annahme, dass die schulische Lern- und Leistungsmotivation durch verschiedene Zielorientierungen bestimmt wird. So unterscheiden Ames und Archer (1988) zwischen Lernzielen und Performanzzielen. Ein wesentliches Unterscheidungsmerkmal ist hierbei die gewählte Bezugsnorm (vgl. Kap. 1.2): Die Lernzielorientierung ist auf die Steigerung eigener Kompetenzen ausgerichtet. Der Vergleich der eigenen Leistung über einen bestimmten Zeitraum hinweg dient dem Lerner als Maßstab zur Bewertung der Kompetenzsteigerung (individuelle Bezugsnorm). Die Performanzzielorientierung dagegen bietet den Vergleich mit anderen Lernern (soziale Bezugsnorm). Lerner mit ausgeprägter Performanzzielorientierung sind weniger an einer Kompetenzsteigerung interessiert als daran, die eigene Kompetenz vor anderen Lernern zu demonstrieren bzw. einen Mangel an Kompetenz zu kaschieren. Der SELLMO spiegelt dieses theoretische Modell der Lern- und Leistungsmotivation durch vier Skalen mit insgesamt 31 Items wider (Tabelle 3).

Untertest	Beschreibung
Leistungsziele	Schüler betrachten Leistung als den Ausdruck eigener Fähigkeiten
Annäherungs-Leistungsziele:	Schüler beabsichtigen, ihre Kompetenzen vor anderen darzustellen
Vermeidungs-Leistungsziele:	Schüler beabsichtigen, Misserfolg vor anderen zu verbergen
Arbeitsvermeidung:	Schüler streben danach, möglichst wenig Anstrengung und Leistung erbringen zu müssen.

Tabelle 3: Untertests des SELLMO

Das Instrument eignet sich in besonderem Maße zur Diagnose motivationaler Defizite bei Schülern, die hinter ihrem tatsächlichen Leistungspotenzial zurückbleiben (Underachiever). Als Konsequenz können gezielte pädagogische Interventionsmaßnahmen eingeleitet werden. Hierzu zählen bspw. die Vermittlung realistischer Zielsetzungen durch die Lehrkräfte sowie die Förderung der individuellen Bezugsnormorientierung. Berger und Rockenbauch (2005) beschreiben das Instrument wir folgt:

> „Die SELLMO-Skalen sind ein methodisch solides und theoretisch fundiertes Instrument, um Minderleistungen bei Schülerinnen und Schülern auf Grund motivationaler Defizite aufzuklären. Hierzu sollten flankierend die intellektuellen Fähigkeiten mittels eines standardisierten Intelligenztests (z. B. HAWIK-III von Wechsler, herausgegeben von Tewes, Rossmann & Schallberger, 2001) sowie das schulische Selbstkonzept mittels des SESSKO (Schöne et al., 2002) abgeklärt werden."

1.4.4 Schulleistungsrelevante Merkmale: Fähigkeitsselbstkonzept

Theoretischer Hintergrund

Die Literatur, die sich mit dem Thema „Selbstkonzept" auseinandersetzt, wird durch eine Vielzahl verschiedener Modelle und Vorstellungen geprägt, was zu einer „babylonischen Sprachverwirrung" führt wie Moschner (2001) anmerkt. Laut Greve (2000) umfasst das Selbstkonzept „alle selbstbezogenen Einschätzungen, Überzeugungen und Meinungen." Die Definition von Shavelson, Hubner und Stanton (1976) basiert auf 7 kennzeichnenden Merkmalen, die wie folgt zusammengefasst werden können:

1. Das Selbstkonzept ist organisiert bzw. strukturiert (d.h. Personen organisieren selbstbezogene, identitätsrelevante Informationen in Kategorien und setzen diese zueinander in Beziehung).
2. Das Selbstkonzept besteht aus verschiedenen Aspekten, welche die Struktur des Selbstkonzepts einer Person widerspiegeln.
3. Das Selbstkonzept ist hierarchisch aufgebaut.

4. Das Selbstkonzept ist abnehmend stabil: die oberen Kategorien der Selbstkonzepthierarchie sind situationsunabhängig (traits), während die unteren Ebenen der Pyramide situationsabhängig sind (states).
5. Das Selbstkonzept ist entwicklungsdynamisch: die verschiedenen Facetten werden mit zunehmendem Alter eindeutiger voneinander abgegrenzt.
6. Das Selbstkonzept hat eine deskriptive (beschreibende) und eine evaluative (bewertende) Komponente.
7. Das Selbstkonzept ist von anderen Konstrukten (z.B. Motivation) unterscheidbar.

Moschner (2001) beschreibt das Selbstkonzept als „ein mentales Modell einer Person über die eigenen Fähigkeiten und Eigenschaften." Nach Shavelson, Huber und Stanton (1976) enthält es neben dem emotionalen, dem sozialen und dem körperlichen Selbstkonzept eine partielle Komponente des generellen Selbstkonzepts. Zur besseren Abgrenzung der verschiedenen Konstrukte steht im vorliegenden Studienbrief lediglich das akademische bzw. schulische Fähigkeitsselbstkonzept im Vordergrund. Das Fähigkeitsselbstkonzept kann definiert werden als „Gesamtheit der kognitiven Repräsentationen eigener Fähigkeiten in akademischen Leistungssituationen" (Dickhäuser, Schöne, Spinath & Stiensmeier-Pelster, 2002).

Selbstkonzept und Leistung

Dass das schulische Fähigkeitsselbstkonzept schulische Leistungen beeinflussen kann, lässt sich aus den verschiedenen Studien schließen, in denen ein signifikanter Zusammenhang zwischen der individuellen Ausprägung des Selbstkonzepts und schulischer Leistung aufgezeigt werden konnte. Das schulische Fähigkeitsselbstkonzept klärt „nennenswerte Beiträge von (Schul-)Leistungsvarianz auf, hängt mit der Ausdauer bei der Bearbeitung von Aufgaben zusammen, beeinflusst das Wahlverhalten (z.B. in der Oberstufe), kovariiert mit Interesse und Leistungsmotivation" (Rost et al., 2007). Die Korrelation der Schulnoten mit dem Selbstkonzept ist teilweise stärker als die Korrelation mit dem IQ (Rost et al., 2007).

Eine Metaanalyse von Hansford und Hattie (1982), die 20 Studien berücksichtigte, ermittelte eine Korrelationsstärke von r = 0,40. Dieser Zusammenhang ist Studien zufolge bereits während der Grundschulzeit evident: Während hochbegabte Schülerinnen und Schüler ein positives Selbstkonzept aufweisen (Rost & Hanses, 1994), konnte bei Schülerinnen und Schülern mit Leistungsschwächen ein negatives Selbstkonzept nachgewiesen werden (Hanses & Rost, 1998).

Als mögliches Erklärungsmodell für diese Befunde wird angenommen, dass das Fähigkeitsselbstkonzept leistungsschwacher Schüler durch den ständigen impliziten oder expliziten Vergleich mit stärkeren Schülerinnen und Schülern im Laufe der Schulzeit weiter absinkt. Diese Hypothese des Bezugsgruppeneffekts wird durch den Befund gestützt, dass Sonderschüler ein höheres Fähigkeitsselbstkonzept als Schülerinnen und Schüler mit gleicher Intelligenz an Regelschulen aufweisen. Der Bezugsnormorientierung (individuell, sozial, kriterial) kommt somit eine entscheidende Rolle bei der relativen Einschätzung eigener Fähigkeiten zu. Für die pädagogische Praxis würde das bedeuten, dass homogene Lerngruppen zu einem positiven Fähigkeitsselbstkonzept beitragen können, da die äquivalente Bezugsgruppe „faire" und realistische Vergleiche ermöglicht. Dennoch muss betont werden, dass die Homogenisierung von Lernenden auch immer ein Prozess der Segregation und Ausgrenzung ist. Individuelle Förderung durch Binnendifferenzierung im Unterricht und durch gezielte Stärkung des Fähigkeitsselbstkonzepts könnten an dieser Stelle als Interventionsmaßnahmen eingesetzt werden. Die von Dickhäuser et al. (2002) vorgeschlagenen Dimensionen des schulischen Fähigkeitsselbstkonzepts finden sich in dem von ihnen entwickelten Instrument wieder, welches im folgenden Abschnitt vorgestellt werden soll.

1.4.4.1 Erfassung des schulischen Selbstkonzepts: Der SESSKO

Die Skalen zur Erfassung des schulischen Selbstkonzepts (SESSKO) umfassen mittels 22 Items vier Untertests (Tabelle 4).

Untertest	Beschreibung
Schulisches Selbstkonzept – kriterial	Einstufung der Leistungen anhand eines sachlichen Kriteriums
Schulisches Selbstkonzept – individuell	Vergleich mit den eigenen Fähigkeiten in der Vergangenheit
Schulisches Selbstkonzept – sozial	Vergleich mit anderen Personen
Schulisches Selbstkonzept – absolut	ohne Vorgabe einer Bezugsnorm erfasst

Tabelle 4: Untertests des SESSKO

Die Aufgaben innerhalb der vier Untertests erfassen gleichermaßen die Bereiche Begabung, Intelligenz, Fähigkeit, Lernfähigkeit und die Bewältigung von Anforderungen. Allerdings wurde der SESSKO dahingehend kritisiert, dass er den Schülerinnen und Schülern eine sehr ausdifferenzierte Selbsteinschätzung in den einzelnen Dimensionen abverlangt, die theoretisch als auch empirisch nicht ausreichend gestützt werden kann.

1.4.5 Weiterführende Literatur

Brähler, E., Holling, H., Leutner, D. & Petermann, F. (Hrsg.). (2002). *Brickenkamp Handbuch psychologischer und pädagogischer Tests*. Band 1 und 2. Göttingen: Hogrefe.

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E.Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule*. Göttingen. Hogrefe.

Hosenfeld, I., & Schrader, F.W. (2006). *Schulische Leistung: Grundlagen, Bedingungen, Perspektiven*. Münster: Waxmann.

1.4.6 Verständnis- und Diskussionspunkte

1. *In welchem Kontext bzw. bei welcher Schülerpopulation wird die Unterscheidung zwischen Fähigkeit und Leistung besonders deutlich?*
2. *Diskutieren Sie die Bedeutung der Aussage, Intelligenz sei etwas „Undefinierbares, aber Messbares."*
3. *Suchen Sie in der UDiKom-Testdatenbank (http://tests.udikom.de) den Culture Fair Test (CFT) und betrachten Sie die Beschreibungen der verschiedenen Untertests. Diskutieren Sie, an welcher Stelle sprachliche oder kulturelle Einflüsse auch beim CFT eine Rolle spielen könnten.*

1.5 Praktische Implikationen

In diesem Kapitel behandelte Fragen:

- *Wie lässt sich die Validität eines Leistungstests gewährleisten?*
- *Wie lässt sich die Objektivität eines Leistungstests bereits durch die Aufgabenkonstruktion erhöhen?*
- *Wie lässt sich die Reliabilität eines Leistungstests mit wenig Aufwand einschätzen?*

Im letzten Kapitel 1.4 wurde der Anwendungsbereich individualdiagnostischer Verfahren skizziert. Wir haben uns also bereits einen Überblick über die Diagnose von Schulleistungsmerkmalen und von schulleistungsrelevanten Merkmalen und deren testtheoretischen Grundlagen verschafft. Im diesem Kapitel sollen nun die praktischen Implikationen der bisher erworbenen diagnostischen Kenntnisse vorgestellt werden. Eine Möglichkeit der praktischen Umsetzung der Wissensinhalte ist die eigenständige Konstruktion eines Testinstruments, um bspw. fachspezifische Kenntnisse und Fähigkeiten zu diagnostizieren. Das Kapitel zur Testtheorie verdeutlichte zwar, dass die Erstellung eines Diagnoseverfahrens mit immensem Aufwand einhergeht, der im Schulalltag kaum zu integrieren ist. Dennoch können die erworbenen Kenntnisse auch im schulischen Alltag genutzt werden. So können bspw. bereits durch die Anwendung bestimmter Konstruktionsprinzipien bei der Entwicklung und Auswahl einzelner Testaufgaben die Voraussetzungen geschaffen werden, dass eigene Klassenarbeiten die klassischen Testgütekriterien angemessen einhalten. Klauer (1987) zeigt bspw. auf, wie durch die Art der Aufgabenkonstruktion und -auswahl die Validität eines Tests (i.S. der Inhaltsvalidität) gewährleistet werden kann, so dass auf eine empirische (und mit hohem Aufwand verbundene) Überprüfung der Validität ggf. verzichtet werden kann. Wir werden diese Vorgehensweise in diesem Kapitel kurz skizzieren.

Validität

Die Objektivität einer Klassenarbeit hängt von der Art der Durchführung, der Auswertung und der Interpretation des Ergebnisses ab. Wie in Kapitel 1.3.4.3 dargelegt, ist eine hohe Objektivität insbesondere durch vorab schriftlich fixierte Vorgaben und Beschreibungen zu erreichen, wie bei der Durchführung, Auswertung und Interpretation vorzugehen ist. Darüber hinaus kann die Objektivität durch eine gute Aufgabenkonstruktion erhöht werden. Wir werden in diesem Kapitel ein paar Hinweise geben, worauf bei der Konstruktion (schriftlicher) Aufgaben geachtet werden sollte.

Objektivität

Die Einschätzung der Reliabilität erfordert die Berechnung eines Korrelationskoeffizienten. Dies ist mit heutiger Standardsoftware für Tabellenkalkulationen wie z.B. Microsoft Excel kein Problem. Auch darauf werden wir im Folgenden kurz eingehen.

Reliabilität

1.5.1 Validität

Im schulischen Kontext ist die Validität eines Leistungstests dann gegeben, wenn der Test misst, ob Schülerinnen und Schüler das gelernt haben, was sie gelehrt bekamen und daher gelernt haben sollen. Insofern ist das Lehrziel, das eine Lehrkraft innerhalb einer Unterrichtseinheit verfolgt, die Prüfgröße für die Validität eines Leistungstests wie z.B. einer Klassenarbeit. Der erste Schritt bei der Konstruktion einer Klassenarbeit ist daher die Analyse dieses Lehrziels und eine Zerlegung des Lehrziels in Teilziele (Klauer, 2001). Die Zerlegung in Teilziele dient der Auflistung aller relevanten Inhaltsbereiche, die durch das Lehrziel angesprochen wurden. Dabei muss diese Liste zum einen vollständig sein, d.h. es darf kein Inhaltsbereich übersehen werden. Zum anderen sollte sie so feingliedrig wie möglich sein. Optimal wäre eine Liste von einzelnen Aussagen, wobei jede Aussage eine Information enthält, die die Schülerinnen und Schüler lernen sollten. Da eine derart feingliedrige Liste auf Aussagenebene jedoch äußerst aufwändig zu erstellen ist, ist eine Liste möglichst kleiner Inhaltsbereiche meist ausreichend und unter Ökonomieaspekten zu bevorzugen.

Zerlegung in Teilziele

	Reproduzieren	Anwenden	Reflektieren	Bewerten
Inhalt A	A	B	C	D
Inhalt B	E	F	G	H
Inhalt C	I	J	K	L

Tabelle 5: Lehrzielmatrix (vgl. Klauer, 2001)

Lehrziel-matrix

Die Inhaltsbereiche lassen sich in die Zeilen einer Tabelle eintragen (Tabelle 5). Die Spalten dieser Tabelle können dann definieren, wie die Schülerinnen und Schüler mit den Inhalten jeweils umgehen können sollen. Reicht ein einfaches Kennen und Reproduzieren der Inhalte, sollen die Inhalte auf neue Gebiete angewandt werden, soll über die Inhalte reflektiert werden oder sollen sie bewertet werden? Welches Verhalten Lernende an den jeweiligen Inhalten zeigen können sollen, ist Teil des Lehrziels und kann sich zwischen Lehrzielen verschiedener Unterrichtseinheiten entsprechend unterscheiden. Wichtig ist jedoch, dass wie bereits die Inhaltsbereiche so auch das gewünschte Verhalten möglichst umfassend beschrieben ist. Denn nur wenn sowohl die Inhaltsbereiche als auch das Verhalten vollständig und umfassend definiert sind und in die Zeilen und Spalten einer Tabelle eingetragen wurden, dann repräsentieren die Zellen der Tabelle (Zellen A – L in Tabelle 5) das Lehrziel vollständig.

Auswahl von Testaufgaben

Die vollständige Repräsentation des Lehrziels ist eine notwendige Voraussetzung für die Inhaltsvalidität eines Tests: Für jede Zelle lassen sich Testaufgaben konstruieren, nach Möglichkeit pro Zelle dieselbe Anzahl von Aufgaben. Aus der so entstandenen Aufgabenmenge werden dann zufällig (oder stratifiziert-zufällig, s. Klauer, 1987) so viele Aufgaben ausgewählt, wie der Test am Ende enthalten soll. Ist jedoch ein Inhaltsbereich in der Tabelle nicht repräsentiert oder ist eine Verhaltensweise nicht in eine der Spalten eingetragen, dann fehlen in der Aufgabenmenge die entsprechenden Aufgaben und der Test verliert seinen Anspruch auf vollständige und umfassende Repräsentation des Lehrziels, sprich seinen Anspruch auf Inhaltsvalidität.

1.5.2 Objektivität

Durchführungs-objektivität

Um die Objektivität eines Leistungstests wie einer Klassenarbeit zu gewährleisten, gilt es zum einen die Durchführungsbedingungen zu standardisieren. Dies ist im schulischen Kontext meist in hohem Ausmaß gegeben: Die Dauer ist durch den Schulstundentakt meist festgelegt, die Aufgabenstellungen werden meist in schriftlicher Form ausgeteilt, die Bedingungen werden für alle Schülerinnen und Schüler gleich gehalten, etc. Aber wie verhält es sich bspw. mit Rückfragen durch einzelne Schülerinnen und Schüler? Welche Fragen werden beantwortet und welche nicht, welche Hilfestellungen werden gegeben? Wenn ein Schüler auf eine Rückfrage einen Hinweis bekommen hat, wird dieser Hinweis dann auch allen weiteren Schülerinnen und Schülern gegeben? Wie verhält es sich mit der Kontrolle? Werden evtl. manche Schülerinnen oder Schüler stärker beobachtet als andere? Vertraut man manchen Schülerinnen und Schülern mehr als anderen? Diese Fragen stehen beispielhaft für mögliche Verletzungen der Durchführungsobjektivität. Sie sollen dazu dienen, den Blick dafür zu schärfen, wie in einer eigentlich recht standardisierten Testsituation das Objektivitätskriterium trotzdem recht leicht verletzt werden kann. Da aufgrund des trotz dieser Gefahren recht hohen Standardisierungsgrades bei Klassenarbeiten die Durchführungsobjektivität jedoch meist sehr akzeptabel ist, soll an dieser Stelle darauf nicht weiter eingegangen werden. Vielmehr wollen wir uns der Auswertungs- und Interpretationsobjektivität zuwenden, die zu einem großen Teil durch das Antwortformat der verwendeten Testaufgaben beeinflusst wird.

1.5.2.1 Schriftliche Testaufgaben mit geschlossenem Antwortformat

Testaufgaben lassen sich zum einen danach unterscheiden, ob ihr Antwortformat schriftlich oder mündlich oder verhaltensbasiert (bspw. im Sport) ist. Wir beschränken uns im Folgenden auf Testaufgaben mit schriftlichem Antwortformat. Hier lässt sich wieder zwischen offenen und geschlossenen Antwortformaten unterscheiden. Bei offenen Antwortformaten wie z.B. kurzen Aufsätzen oder Portfolios kann die richtige oder optimale Lösung der Aufgabe in vielen verschiedenen Varianten niedergeschrieben werden bzw. es kann mehrere richtige oder optimale Lösungen der Aufgabe geben. Bei geschlossenen Aufgabenformaten wie z.B. Ergänzungsaufgaben oder Mehrfach-Wahlaufgaben (Multiple-Choice-Aufgaben) gibt es genau eine vorab bestimmte richtige Lösung, die sich durch ein genau definiertes Wort oder auch nur ein richtig gesetztes Kreuz o.ä. ausdrückt. Damit erlauben geschlossene Antwortformate deutlich weniger Spielraum bei der Auswertung des Tests, wodurch die Objektivität von Testaufgaben mit geschlossenem Antwortformat als deutlich höher einzuschätzen ist als bei Testaufgaben mit offenem Aufgabenformat.

Doch der Vorteil der höheren Objektivität von Testaufgaben mit geschlossenem Aufgabenformat geht einher mit dem Nachteil, dass diese Art der Testaufgaben meist schwieriger zu konstruieren ist. Zudem haben Testaufgaben mit geschlossenem Aufgabenformat den Ruf, nur kognitiv wenig anspruchsvolle Fähigkeiten wie den reinen Abruf von Wissen testen zu können. Dass dieser Ruf jedoch nicht gerechtfertigt ist, demonstriert u.a. Klauer (2001) eindrucksvoll anhand von Mehrfach-Wahlaufgaben aus der TIMS-Studie oder dem Mediziner-Test. Doch wenn Testaufgaben mit geschlossenem Antwortformat besser sein sollen als ihr Ruf, dann müssen bei der Aufgabenkonstruktion einige Faustregeln beachtet werden, was mit einem gewissen Aufwand verbunden ist. Diese Faustregeln sollen im Folgenden für die gängigsten geschlossenen Antwortformate besprochen werden.

Ergänzungsaufgaben

Ergänzungsaufgaben eignen sich nicht nur für die Wissensabfrage, sondern auch für (wenig komplexe) Formen der Wissensanwendung (bspw. das Lösen von Bruchrechenaufgaben) und des Verständnisses (bspw. Steigerung eines Adjektivs). Einige Beispiele (gut sowie schlecht konstruierter) Ergänzungsaufgaben finden sich in Tabelle 6.

Ergänzungsaufgaben werden konstruiert, indem aus einer Aussage ein Wort oder eine Zahl entfernt und durch eine Leerstelle ersetzt wird. Wurde das Lehrziel bis auf Aussagenebene in Teilziele zerlegt, können diese Aussagen herangezogen werden. Andernfalls müssen Aussagen konstruiert werden, die jeweils ein Teilziel repräsentieren. Wichtig dabei ist, dass die Aussagen keine wörtlichen Zitate aus Lehrtexten darstellen, da dieses ein reines Auswendiglernen seitens der Schülerinnen und Schüler befördert. Durch eine entsprechende klare Instruktion werden die Schülerinnen und Schüler dann gebeten, in die Leerstelle das fehlende Wort bzw. die fehlende Zahl einzutragen.

Faustregeln

Die Aussagen sollten so konstruiert sein, dass die Leerstellen möglichst weit am Ende des Satzes stehen. Das ermöglicht den Schülerinnen und Schülern, möglichst viele Informationen zunächst zu lesen, bevor sie eine Antwort finden müssen. Um keine Hinweise auf das Lösungswort zu geben, sollten alle Leerstellen im Text dieselbe Länge aufweisen. Zudem sollte vermieden werden, dass grammatikalische Hinweise wie z.B. Artikel manche (falsche) Lösungswörter ausschließen. Wichtig ist, dass es für jede Leerstelle genau ein einziges richtiges Lösungswort bzw. genau eine richtige Zahl gibt. Genauso sollte in einer Aussage möglichst nur eine einzige Leerstelle gesetzt werden, da ansonsten die Gefahr besteht, dass ein falsches Ausfüllen der einen Leerstelle auch zu einem fehlerhaften Ausfüllen der weiteren Leerstelle führt, sprich die Leerstellen nicht unabhängig voneinander ausgefüllt werden können.

Ergänze jede Leerstelle so, dass die Aussage stimmt. Schreibe dabei deutlich und richtig. Jede richtige Antwort gibt einen Punkt.	
Die Evolutionstheorie von _____ basiert auf dem Prinzip der _____.	Mehrere Leerstellen; Grammatikalischer Hinweis
Columbus entdeckte Amerika _____.	Mehrere mögliche richtige Antworten
16 + 7 * 2 = ____.	gut
In welchem Jahr wurde Helmut Kohl zum ersten Mal Bundeskanzler der Bundesrepublik Deutschland? _____	gut
Zu ____ Leerstellen frustrieren sowohl _____ als auch _____.	Mehrere Leerstellen; Leerstellen unterschiedlicher Länge; Leerstelle am Beginn der Aussage

Tabelle 6: Beispiele für Ergänzungsaufgaben inklusive Bewertungen

Zuordnungsaufgaben

Zuordnungsaufgaben bestehen aus einer Liste von Aussagen und einer Liste von Optionen. Die Aufgabe besteht darin, jeder Aussage genau eine der Optionen zuzuordnen. Zwei Beispiele für Zuordnungsaufgaben sind in Tabelle 7 dargestellt.

Aufgabe 1:		
Ordne jeder Persönlichkeit eine Aussage zu.		
1. Gott ist tot	1. Marx	Keine Homogenität
2. Gott als Projektion	2. Freud	Hohe Ratewahrscheinlichkeit
3. Religion als Opium des Volkes	3. Nietzsche	Keine Ordnung
4. Religion als kollektive Zwangsneurose	4. Lennon	Keine Eindeutigkeit
5. Gott ist ein Konzept	5. Feuerbach	Mangelhafte Instruktion
Aufgabe 2:		
In der linken Spalte steht, was eine Person erfunden hat, in der rechten Spalte stehen berühmte Erfinder. Ordne den Erfindungen ihren Erfinder zu, indem du den entsprechenden Buchstaben auf die Linie vor der Erfindung schreibst.		
___1. Er hat die Entkörnungsmaschine für Baumwolle erfunden.	a. Alexander G. Bell	
___2. Eine seiner Erfindungen war das Telefon.	b. Henry Bessemer	
___3. Er hat das Radio erfunden.	c. Thomas Edison	
	d. Guglielmo Marconi	gut
	e. Eli Whitney	
	f. Orville Wright	

Tabelle 7: Beispiele für Zuordnungsaufgaben inklusive Bewertungen

Faustregeln

Die Instruktionen enthalten die Angabe, ob jede Option genau einmal oder mehrfach zugeordnet werden muss und auf welche Weise zugeordnet werden soll. Die Aussagen sowie die Optionen stammen aus homogenen Inhaltslisten (nicht wie im oberen Beispiel in Tabelle 7, wo in den Optionen Politiker und Komiker vermischt werden), damit einzelne Optionen nicht mehr ins Auge stechen als andere. Es sollten nicht mehr als zehn Aussagen bzw. Optionen pro Aufgabe verwendet werden, wobei die Anzahl der Optionen größer sein sollte als die Anzahl der Aussagen, was die Ratewahrscheinlichkeit bei der Zuordnung verringert. Jeder Aussage sollte genau eine richtige Option zugeordnet werden können, wenngleich Mehrfachzuordnungen durchaus möglich sind. Darauf muss dann in der Instruktion jedoch explizit hingewiesen werden. Der Übersichtlichkeit wegen sollte eine Aufgabe inklusive der Aussagen und Optionen nicht über zwei Seiten hinweg präsentiert werden. Aussagen und Optionen sollten nummeriert sein, jedoch mit unterschiedlichen Nummerierungen (bspw. Nummern für die Aussagen und Buchstaben für die Optionen).

Wahr-/Falsch-Aufgaben

Wahr-/Falsch-Aufgaben ermöglichen eine effiziente Abdeckung umfangreicher Inhaltsbereiche, insbesondere wenn das Lehrziel und seine Teilziele bereits auf Aussagenebene definiert sind. Der Aufwand bei der Konstruktion und Auswertung der Aufgaben ist verglichen mit alternativen Antwortformaten gering. Werden bei der Konstruktion von Wahr-/Falsch-Aufgaben jedoch bestimmte Gestaltungsregeln außer Acht gelassen, reduziert sich der diagnostische Wert dieser Aufgaben auf eine reine Abfrage trivialen Faktenwissens, die mit sehr hoher Ratewahrscheinlichkeit einhergeht und Schülerinnen und Schüler zur unreflektierten Akzeptanz vereinfachter Aussagen verleiten kann. Beispiele für Wahr-/Falsch-Aufgaben sind in Tabelle 8 dargestellt.

Gib für jede Aussage an, ob sie wahr oder falsch ist.	wahr	falsch	
Der Monoghalea fließt nach Norden, wo er sich bei Columbus mit dem Allegheny vereint und damit den Ohio bildet.	□	□	Mehrere Aussagen
Lange Tests sind immer reliabler als kurze Tests.	□	□	absolute Formulierung
Lange Tests sind meistens reliabler als kurze Tests.	□	□	abschwächende Formulierung
Gebete sollten in der Schule verboten sein.	□	□	Meinung ohne konkreten Bezug
5 + 3 * 2 = 16	□	□	gut
Wenn ein Flugzeug genau auf der Grenze zwischen Deutschland und Frankreich abstürzt, wird die eine Hälfte der Überlebenden in Deutschland und die andere Hälfte in Frankreich beigesetzt.	□	□	Fangfrage

Tabelle 8: Beispiele für Wahr-/Falsch-Aufgaben inklusive Bewertungen

Faustregeln

Bei der Konstruktion von Wahr-/Falsch-Aufgaben sollte in einer Aufgabe genau eine Aussage repräsentiert sein (und nicht mehrere Aussagen miteinander verknüpft werden wie im ersten Beispiel in Tabelle 8) und diese sollte zweifelsfrei und ohne weitere Erläuterung als wahr oder falsch bewertbar sein. Die Aufgaben sollten auf eine Wissens- und Verständnisabfrage abzielen und keine Meinungen erfragen, die in diesem Antwortformat nicht begründbar sind. Die Aussagen sollten auch typische Fehlkonzepte repräsentieren, die dann als falsch zu bewerten wären. Bei der Konstruktion von Aufgaben, die als wahr zu bewerten sind, tendiert man häufig dazu, die Aussage sehr präzise zu formulieren und damit die Aussagenlänge zu erhöhen. Diese Tendenz ist bei Aussagen, die als falsch zu bewerten sind, nicht so stark ausgeprägt, wodurch es leicht passiert, dass Wahr-Aussagen lang und Falsch-Aussagen kurz formuliert sind. Ein solcher systematischer Unterschied in der Aussagenlänge sollte unbedingt vermieden werden. Ebenso zu vermeiden sind doppelte Verneinungen, die schnell überlesen werden, oder auch absolute oder abschwächende Formulierungen, da absolut formulierte Aussagen mit hoher Wahrscheinlichkeit falsch sind, Aussagen mit abschwächenden Aussagen mit hoher Wahrscheinlichkeit wahr. Testerfahrene Schülerinnen und Schüler können davon profitieren ohne Kenntnisse und Fähigkeiten im eigentlich zu testenden Inhaltsbereich zu haben. Wie für alle Aufgaben mit geschlossenem Aufgabenformat gilt auch für Wahr-/Falsch-Aufgaben, dass die Aussagen keine direkten Zitate aus Lehrbüchern sein sollten (um ein Auswendiglernen zu vermeiden), dass keine Fangfragen zu verwenden sind und dass bei mehreren zu bewertenden Aussagen keine Systematik in der Reihenfolge von wahren und falschen Aussagen erkennbar sein sollte.

Mehrfach-Wahlaufgaben (Multiple Choice)

Mehrfach-Wahlaufgaben (Multiple Choice-Aufgaben, MC-Aufgaben) bieten nicht nur die Möglichkeit einer vertiefenden Wissensabfrage, sondern auch der Erfassung von Verständnis-, Anwendungs- und Transferleistungen. Allerdings ist dafür die Konstruktion von MC-Aufgaben äußerst aufwändig. Eine MC-Aufgabe setzt sich aus einem Aufgabenstamm inklusive der Fragestellung, einer Anweisung und den Optionen zusammen (Tabelle 9).

Faustregeln Aufgabenstamm

Der Aufgabenstamm wird vollständig vor den Alternativen präsentiert. Er enthält alle notwendigen Informationen, die die Schülerinnen und Schüler für das Verständnis der Aufgabe benötigen, und nicht mehr. Auf irrelevante Ausschmückungen sollte verzichtet werden und das Vokabular sowie die Satzstruktur sollten möglichst einfach sein, um den Leseaufwand nicht unnötig zu erhöhen. Negative Formulierungen sollten vermieden werden. Zudem sollten auch bei MC-Aufgaben direkte Zitate aus Lehrbüchern vermieden werden. Der Aufgabenstamm endet mit einer direkten Frage, zu der die folgenden Optionen jeweils eine mögliche Antwort darstellen. Hierbei sollte überprüft werden, ob die Frage womöglich grammatikalische Hinweise auf die richtige Option (den Attraktor) enthält. Auf eine Abfrage persönlicher Meinungen sollte verzichtet werden.

Die Anweisung enthält die Angabe, ob genau eine Option richtig ist oder ob keine oder mehrere Optionen richtig sein können.

Faustregeln Optionen

Die Konstruktion der Optionen ist das Kernstück der MC-Aufgabenkonstruktion. Üblicherweise werden pro Aufgabe 3 bis 5 Optionen gegeben, die aus einer homogenen Inhaltsliste stammen und ungefähr dieselbe Länge haben sollten. Sollten alle Optionen mit denselben Worten beginnen, können diese über die Optionen geschrieben werden und mit […] mit den Optionen verbunden werden, um den Leseaufwand zu reduzieren. Auch bei der Formulierung der Optionen gilt, dass keine direkten Zitate aus Lehrbüchern und keine absoluten oder abschwächenden Formulierungen verwendet werden sollten. Die Optionen

müssen unabhängig voneinander bewertbar sein, was auch für Optionen unterschiedlicher Aufgaben desselben Tests gilt.

Bei den Optionen unterscheidet man zwischen dem Attraktor und den Distraktoren. Der Attraktor ist die richtige Option, die Distraktoren repräsentieren falsche Antworten auf die im Aufgabenstamm präsentierte Frage. Der Attraktor muss ohne weitere Erläuterung als richtig bewertbar sein. Bei mehreren MC-Aufgaben in einem Test sollte darauf geachtet werden, dass die Position der Attraktoren keine Systematik aufweist (z.B. immer die erste Option als Attraktor). Die Konstruktion der Distraktoren ist die größte Herausforderung. Sie haben die Funktion, Personen mit geringen Kenntnissen und Fähigkeiten von dem Attraktor abzulenken. Dafür müssen sie (für diese Personen) plausibel sein. Daher bieten sich insbesondere typische Fehlkonzepte als Grundlage für die Distraktorenkonstruktion an. Je ähnlicher Distraktoren zum Attraktor sind, desto schwieriger machen sie die Aufgabe.

Welcher Autor schrieb den Montageroman „Die neuen Leiden des jungen W.“?	Aufgabenstamm
Kreuzen Sie die eine richtige Antwort an.	Anweisung
	Optionen:
☐ Büchner	Distraktor
☐ Fontane	Distraktor
☐ Goethe	Lockvogel
☐ Plenzdorf	Attraktor
☐ Schiller	Distraktor

Tabelle 9: Beispiele für eine Mehrfach-Wahlaufgaben

1.5.2.2 Schriftliche Testaufgaben mit offenem Antwortformat

Aufgaben mit einem offenen Antwortformat wie z.B. Aufsätze bieten den Vorteil, bei relativ geringem Konstruktionsaufwand relativ komplexe Inhaltsbereiche und Leistungen adressieren zu können. Der geringe Aufwand bei der Konstruktion der Aufgaben wird jedoch erkauft durch einen recht hohen Aufwand bei der Auswertung (und Interpretation) der Antworten sowie durch eine häufig geringe Objektivität. Auch wenn dieses Problem wohl nie vollständig gelöst werden kann, so gibt es doch einige Maßnahmen, um zumindest die Objektivität dieser Aufgaben zu erhöhen. Diese Maßnahmen betreffen zum einen die Aufgabenstellung selbst sowie das Vorgehen bei der Auswertung.

Faustregeln Aufgabenstellung

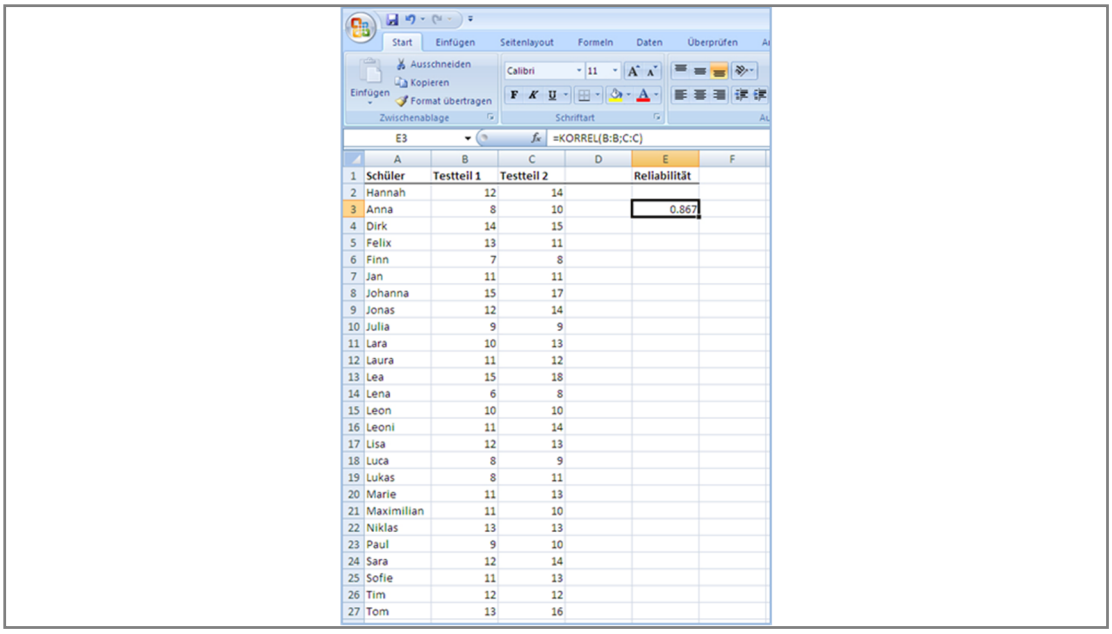
Bevor eine Aufgabenstellung formuliert wird, sollte man sich darüber bewusst sein, welche Art von Verhalten von den Schülerinnen und Schülern erwartet wird. Wenn „nur“ ein einfaches Reproduzieren oder Anwenden verlangt ist, dann lässt sich dieses auch mit Aufgaben mit geschlossenem Antwortformat erfassen. Nur wenn die Anforderungen in Richtung Bewerten oder Reflektieren gehen, ein Anforderungsniveau, für das die Konstruktion von Aufgaben mit geschlossenem Antwortformat sehr aufwändig ist, dann lohnt es sich, das offene Aufgabenformat und den damit verbundenen Auswertungs- und Interpretationsaufwand zu wählen. Die Aufgabenstellung sollte in dem Fall klar formuliert sein und keine Was-/Wer-/Wann-Fragen enthalten, die eine reine Wissensreproduktion verlangen. Sie sollte Angaben über den erwarteten Inhalt und das erwartete Verhalten/das erwartete Niveau enthalten genauso wie Angaben über evtl. Seiten- oder Zeitbeschränkungen, über die Bewertung der Organisation und Struktur des Textes sowie über den Umgang mit Rechtschreib- und Grammatikfehlern. Wenn die Aufgabenstellung das Darlegen eines Standpunktes bzw. einer Meinung verlangt, dann sollte deutlich gemacht werden, dass nicht die Meinung selbst, sondern die Begründung der Meinung bewertet wird. Aus Gründen der Vergleichbarkeit sollte auf verschiedene zur Wahl stehende Aufgabenstellungen verzichtet werden.

Faustregeln Auswertung

Für die Auswertung sollte aus Gründen der Objektivität auf ein vorab erstelltes Bewertungsschema zurückgegriffen werden, das bspw. in Form einer Checkliste formuliert sein kann. Dieses Bewertungsschema enthält die Kriterien, nach denen die offenen Antworten zu bewerten sind (und die in der Aufgabenstellung den Schülerinnen und Schülern auch genannt wurden). Diese Kriterien werden zudem in ihrer Ausprägung beschrieben, die erreicht werden muss, um für dieses Kriterium die volle Punktzahl zu erhalten. Möglich und gängig sind auch Einschätzungen der Qualität einzelner Kriterien auf einer sog. Likert-Skala. Dabei werden verschiedene Aussagen dahingehend bewertet, inwiefern sie für den Aufsatz zutreffen. Bspw. könnte die Aussage „In dem Aufsatz werden die relevanten Pro- und Contra-Argumente klar verständlich erläutert“ dahingehend bewertet werden ob sie (1) „nicht zutrifft“, (2) „eher nicht zutrifft“, (3) „weder zutrifft noch nicht zutrifft“, (4) „eher zutrifft“ oder (5) „zutrifft“.

1.5.3 Reliabilität

Um die Reliabilität eines Tests zu erhöhen und sie zudem überprüfbar zu machen, gibt es eine goldene Regel: Je mehr Aufgaben ein Test zur Erfassung einer Leistung bzw. Fähigkeit enthält, desto besser. Ein Testergebnis, das als Summe oder Mittelwert über viele Testaufgaben berechnet wird, hat mit hoher Wahrscheinlichkeit eine höhere Reliabilität als ein Ergebnis, das nur auf sehr wenigen Testaufgaben, sprich auf sehr wenigen Messungen beruht. Je weniger Messungen, desto stärker fallen die Fehler einer Messung ins Gewicht. Um also die Reliabilität eines Tests zu erhöhen, sollte man versuchen, Tests und Klassenarbeiten zu konstruieren, bei denen dieselbe Fähigkeit wiederholt durch mehrere/viele Aufgaben getestet wird.



	A	B	C	D	E	F
1	Schüler	Testteil 1	Testteil 2		Reliabilität	
2	Hannah	12	14			
3	Anna	8	10		0,867	
4	Dirk	14	15			
5	Felix	13	11			
6	Finn	7	8			
7	Jan	11	11			
8	Johanna	15	17			
9	Jonas	12	14			
10	Julia	9	9			
11	Lara	10	13			
12	Laura	11	12			
13	Lea	15	18			
14	Lena	6	8			
15	Leon	10	10			
16	Leoni	11	14			
17	Lisa	12	13			
18	Luca	8	9			
19	Lukas	8	11			
20	Marie	11	13			
21	Maximilian	11	10			
22	Niklas	13	13			
23	Paul	9	10			
24	Sara	12	14			
25	Sofie	11	13			
26	Tim	12	12			
27	Tom	13	16			

Abbildung 12: Reliabilitätsberechnung mit Excel

Berechnung der Reliabilität

Eine Mehrzahl von Aufgaben ist gleichzeitig die notwendige Voraussetzung für die Überprüfung der Reliabilität. Wie in Kapitel 1.3.4.1 dargelegt, erfolgt eine Reliabilitätsprüfung prinzipiell durch die Berechnung eines Korrelationskoeffizienten, wofür für jede getestete Person mindestens zwei Messungen vorliegen müssen. Die Berechnung einer Korrelation ist mit gängiger Tabellenkalkulationssoftware wie Microsoft Excel einfach zu realisieren. Alles, was benötigt wird, sind zwei Messungen (zwei Zahlen) für jede getestete Person, die man in ein Tabellenblatt einträgt (Abbildung 7). Diese zwei Zahlen erhält man bspw., wenn man nach der split-half- oder der odds/even-Methode (s. Kap. 1.3.4.1) die Aufgaben eines Tests in zwei Gruppen aufteilt und für jede Person eine Summe oder einen Mittelwert für diese beiden Aufgabengruppen (=Testteile) berechnet. Diese Werte sind in das Tabellenblatt so einzutragen, dass in einer Spalte (in Abbildung 12 ist das Spalte B) die Werte für den einen Testteil und in einer zweiten Spalte (Spalte C in Abbildung 12) die Werte für den anderen Testteil stehen. Excel berechnet die Korrelation der beiden Testteile, wenn man in eine leere Zelle den Befehl „=KORREL(B:B;C:C)“ schreibt, wobei „B:B“ für die Spalte mit den Werten des ersten Testteils steht und „C:C“ für die Spalte mit den Werten des zweiten Testteils. Ein Korrelationskoeffizient von $r = 0{,}80$ oder höher ist sehr zufriedenstellend. Für die schulische Praxis sind aber auch Koeffizienten größer $r = 0{,}60$ durchaus zufriedenstellend.

1.5.4 Weiterführende Literatur

Hanna, G.S. & Dettmer, P.A. (2004). *Assessment of effective teaching. Using context-adaptive planning*. Boston: Pearson.

Kubiszyn, T. & Borich, G. (2003). *Educational testing and measurement. Classroom application and practice*. (7th ed.). New York: Wiley.

Nitko, A.J. (2004). *Educational assessment of students*. (4th ed.). Upper Saddle River, NJ: Pearson.

1.6 Literatur

Amelang, M., &. Schmidt-Atzert L. (2006). *Psychologische Diagnostik und Intervention*. Heidelberg: Springer.

Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology, 80,* 260-267.

Berger, U., & Rockenbauch, K. (2005). Testbesprechung: Skalen zur Erfassung der Lern- und Leistungsmotivation (SELLMO). *Diagnostica, 51*, 207-211.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. Heidelberg: Springer.

Bortz, J., & Döring, N. (2006). Forschungsmethoden *und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer.

Cattell, R. B. (1968). Are IQ-Tests intelligent? *Psychology Today*, *2*, 56-62.

Cattell, R. B., & Piaggio, L. (1973). *Die empirische Erforschung der Persönlichkeit*. Weinheim: Beltz.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Degen, R. (2000). *Lexikon der Psychoirrtümer. Warum der Mensch sich nicht therapieren, erziehen und beeinflussen lässt*. Frankfurt: Eichborn.

Dickhäuser, O. Schöne, C., Spinath, B. und Stiensmeier-Pelster, J. (2002). Skalen zum akademischen Selbstkonzept: Konstruktion und Überprüfung eines neuen Instruments. *Zeitschrift für Differentielle und Diagnostische Psychologie, 23*, 393-405.

Greve, W. (2000). Die Psychologie des Selbst: Konturen eines Forschungsthemas. In W. Greve (Hrsg.), *Die Psychologie des Selbst* (S. 15-36). Weinheim: PVU.

Gröschke, D. (2005). *Psychologische Grundlagen für Sozial- und Heilpädagogik. Ein Lehrbuch zur Orientierung für Heil-, Sonder- und Sozialpädagogen*. Bad Heilbrunn: Klinkhardt.

Hanses, P. & Rost, D. H. (1998). Das "Drama" der hochbegabten Underachiever –„Gewöhnliche" oder „außergewöhnliche" Underachiever? *Zeitschrift für Pädagogische Psychologie, 12*, 53–71.

Hansford, B. C. & Hattie, J. A. (1982). The relationship between self and achievement/performance measures. *Review of Educational Research, 52*, 123-142.

Heller, K. A. (1984). Schulleistungsdiagnostik: Einleitung und Übersichtsreferat. In K. A. Heller (Hrsg.), *Leistungsdiagnostik in der Schule* (S. 15-38). Bern: Huber.

Hosenfeld, I., & Schrader, F. W. (2006). *Schulische Leistung: Grundlagen, Bedingungen, Perspektiven*. Münster: Waxmann.

Ingenkamp, K., & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6., neu ausgestattete Aufl.). Beltz Pädagogik. Weinheim: Beltz.

Klauer, K. J. (1987). *Kriteriumsorientierte Tests*. Göttingen: Hogrefe.

Klauer, K. J. (2001). Wie misst man Schulleistungen?. In F. E. Weinert (Hg.), *Leistungsmessungen in Schulen* (S. 103-115). Weinheim: Beltz .

Kliemann, S. (Hrsg.) (2008). Diagnostizieren und Fördern in der Sekundarstufe I. Schülerkompetenzen erkennen, unterstützen und ausbauen. Berlin: Cornelsen Scriptor.

Langfeldt, H.-P. (1984). Die Klassische Testtheorie als Grundlage normorientierter (standardisierter) Schulleistungstests. In: K. A. Heller (Hrsg.), *Leistungsdiagnostik in der Schule* (S. 65-98). Bern: Huber.

Langfeldt, H.-P. (2006). *Psychologie für die Schule*. Weinheim: Beltz.

Lienert, G. A. (1961). *Testaufbau und Testanalyse*. Weinheim: Beltz.

Moosbrugger, H., & Kelava, A. (Hrsg.). (2007). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer.

Moschner, B. (2001). Selbstkonzept. In: D. Rost (2001). *Handwörterbuch pädagogische Psychologie*, Weinheim: Beltz.

Rheinberg, F. (1980). *Leistungsbewertung und Lernmotivation*. Göttingen: Hogrefe.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsmessung. In: F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 59-71). Weinheim: Beltz.

Rindermann, H. & Kwiatowski, (2010). Diagnostik von Intelligenz. In C. Quaiser-Pohl & H. Rindermann (Hrsg.), *Entwicklungsdiagnostik*. München: Reinhardt.

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

Rost, D.H. & Hanses, P. (1994). Besonders begabt: besonders glücklich, besonders zufrieden? Zum Selbstkonzept hoch- und durchschnittlich begabter Kinder. *Zeitschrift für Psychologie, 202*, 379-403.

Rost, D. H., Sparfeldt, J. & Schilling, S. R. (2007). *DISK-Gitter mit SKSLF-8. Differentielles Schulisches Selbstkonzept-Gitter mit Skala zur Erfassung des Selbstkonzepts schulischer Leistungen und Fähigkeiten.* Göttingen: Hogrefe.

Schrader, F.-W. & Helmke, A. (2001). Alltägliche Leistungsbeurteilung durch Lehrer. In F. E. Weinert (Hrsg.) *Leistungsmessungen in Schulen* (S. 45-58). Weinheim: Beltz.

Shavelson, R.J., Huber, J.J. & Stanton, G.C. (1976). *Self-concept: Validation of construct interpretations*. Review of Educational Research, 46, pp. 407-441.

Wild, E., Hofer, M., & Pekrun, R. (2001). Psychologie des Lerners. In A. Krapp, B. Weidenmann (Hrsg.), *Pädagogische Psychologie. Ein Lehrbuch* (S. 207-270). Weinheim: Beltz PVU.

Vergleichsarbeiten

Christian Spoden
Detlev Leutner

UDiKom

Aus- und Fortbildung der Lehrkräfte in Hinblick auf Verbesserung der Diagnosefähigkeit, Umgang mit Heterogenität, individuelle Förderung

Vergleichsarbeiten

Alle im Projekt erstellten Materialien
finden Sie unter
www.udikom.de

2.1 Gegenstand und Zielsetzungen

Vergleichsarbeiten stellen ein relativ junges Instrument der pädagogisch-psychologischen Diagnostik dar, um dessen Zielsetzungen es in der Lehrerschaft große Unsicherheit gibt. Im folgenden Kapitel wird zunächst der Entwicklungshintergrund von Vergleichsarbeiten vorgestellt. Im Anschluss werden dann der Gegenstandsbereich in Abgrenzung zur Individualdiagnostik und dem Bildungsmonitoring sowie die auf die Nutzung für schulinterne Evaluationszwecke ausgerichteten Zielsetzungen erläutert.

In diesem Kapitel werden folgende Fragen beantwortet:

- *Wieso wurden Vergleichsarbeiten als dritte Säule der Qualitätssicherung in Deutschland implementiert?*
- *Mit welchen Zielsetzungen sind Vergleichsarbeiten ausgestattet und wie lassen sich diese von Zielsetzungen der Individualdiagnostik und dem Bildungsmonitoring abgrenzen?*
- *Durch welche charakteristischen Merkmale lassen sich Vergleichsarbeiten von Individualdiagnostik und Bildungsmonitoring abgrenzen?*

2.1.1 Hintergrund

Hintergrund

Das erwartungswidrig schwache Abschneiden deutscher Schülerinnen und Schüler im Vergleich der OECD-Staaten bei den internationalen Schulleistungsvergleichsstudien TIMSS II (Third International Mathematics and Science Study; Baumert et al., 1997) und PISA 2000 (Programme for International Student Assessment; Baumert et al., 2001; Baumert et al., 2002; vgl. Studienbrief 3) veranlasste die deutsche Kultusministerkonferenz (KMK) zur Vorbereitung weitreichender Veränderungen im deutschen Bildungssystem. Diese sollten die Qualitätssicherung von Unterricht und Schule auf Basis der *Setzung*, *Normierung und Überprüfung von Bildungsstandards* in den Vordergrund rücken. In Kooperation mit dem neu gegründeten „Institut zur Qualitätsentwicklung im Bildungswesen“ (IQB; http://www.iqb.hu-berlin.de/) formulierte die KMK daher eine Gesamtstrategie zum Bildungsmonitoring, die einen „Dreiklang aus mehr Eigenständigkeit für Schulen bei gleichzeitiger Vorgabe verbindlicher Standards und bei regelmäßiger Evaluation“ betonte (KMK & IQB, 2006). Als Reaktion auf den „PISA-Schock“ wurden durch die Bundesländer aber auch eigenständig zahlreiche Maßnahmen ergriffen, die zur Weiterentwicklung und Qualitätssicherung des Unterrichts in den Schulen führen sollten. Darunter lässt sich im Rahmen der *Standardsetzung* die an den nationalen Bildungsstandards orientierte und diese konkretisierende Überarbeitung der Lehrpläne fassen. Als Maßnahme der *Standardüberprüfung* führten nahezu alle Bundesländer außerdem flächendeckende Vergleichsarbeiten (in einigen Bundesländern unter den Namen *Diagnosearbeiten, Jahrgangsstufentest, Lernstandserhebungen, Kompetenztests oder Orientierungsarbeiten*) in der Grundschule und Sekundarstufe I ein. Die KMK griff nun wiederum ihrerseits diese Vorarbeiten der Länder auf und beschloss, regelmäßig durchgeführte Vergleichsarbeiten neben den internationalen Schulleistungsstudien, dem Vergleich der Bundesländer und der gemeinsamen Bildungsberichterstattung von Bund und Ländern in die oben skizzierte Gesamtstrategie einzubinden (KMK, 2006; vgl. auch Abb. 1).

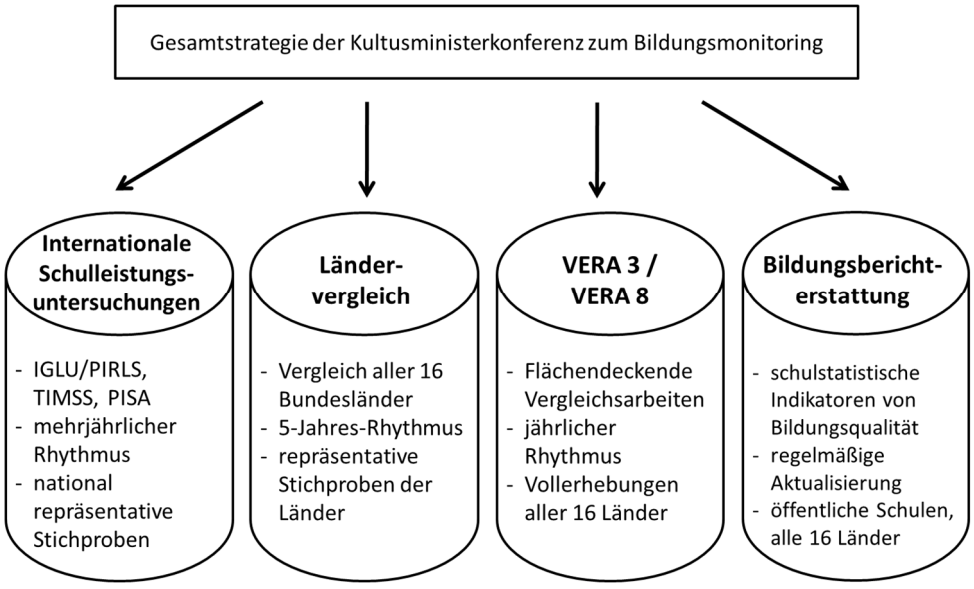


Abbildung 1: Vier Säulen der Qualitätssicherung in deutschen Schulen entsprechend der Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring (KMK, 2006)

Adressaten

Adressaten der diagnostischen Informationen aus Vergleichsarbeiten sind in vielen Bundesländern zunächst die Fachkonferenzen innerhalb der Schulen. Ihnen wird im Rahmen einer objektiven Leistungsrückmeldung Orientierung gegeben, inwieweit die erwarteten Standards innerhalb einer Klasse und innerhalb einer Jahrgangsstufe erreicht werden konnten. Diese Rückmeldung über Schülerleistungen ist mit der Erwartung verbunden, pädagogische, didaktische und ggf. curriculare Veränderungen in Gang zu setzen. Vergleichsarbeiten dienen der „landesweiten, jahrgangsbezogenen Untersuchung des Leistungsstands aller Schulen und Klassen" (KMK & IQB, 2006, S. 21) im Hinblick auf die länderübergreifenden Bildungsstandards. Es verbleibt in der Verantwortung der Fachkonferenzen, im Anschluss an die Ergebnisrückmeldung aus den Arbeiten Konsequenzen für das Lehren und Lernen in der Schule abzuleiten (vgl. hierzu Kapitel 2.5). Hieraus erwächst die pädagogische Herausforderung der Vergleichsarbeiten für Lehrkräfte.

2.1.2 Gegenstand und Charakteristika von Vergleichsarbeiten

Merkmale von Vergleichsarbeiten

Vergleichsarbeiten[1] sind schriftliche Arbeiten der Schülerinnen und Schüler, die auf Basis vorgegebener Aufgabenstichproben landesweit in Teilleistungsbereichen ausgewählter Kernfächer mit dem Ziel durchgeführt werden, Schulleistungen orientiert an einer kriterialen und sozialen Bezugsnorm zu erfassen (vgl. Helmke & Hosenfeld, 2003a). Die Arbeiten finden in den Fächern Deutsch und Mathematik in der Jahrgangstufe 3 sowie den Fächern Deutsch, Mathematik und der ersten Fremdsprache (Englisch/Französisch) in der Jahrgangstufe 8 an einheitlichen Tagen und unter standardisierten Bedingungen statt.[2] Die Kompetenzerwartungen sind durch die nationalen Bildungsstandards für die Grundschule (KMK, 2004a-b) bzw. den mittleren Bildungsabschluss und Hauptschulabschluss (KMK, 2003a-c; 2004c-e) vorgegeben, welchen die Lehrpläne der Bundesländer verpflichtet sind. Aus den Bildungsstandards werden in jedem Jahr für die Vergleichsarbeiten bestimmte Kompetenz- oder Inhaltsbereiche ausgewählt.

Im Gegensatz zu Klassenarbeiten, welche das unmittelbar vorangegangene Unterrichtsgeschehen thematisieren, beziehen sich *Vergleichsarbeiten* auf die Erfassung von Kompetenzen (Fertigkeiten, Fähigkeiten, Kenntnisse; vgl. die folgende Erläuterung) bis zu einem festgelegten Zeitpunkt in der Bildungsbiographie. Als ein weiteres Merkmal von Vergleichsarbeiten sei die Einhaltung von Testgütekriterien genannt. Die Tests sind dazu im Vorfeld umfangreich erprobt worden (vgl. Kapitel 2.3.3; http://www.iqb.hu-berlin.de/vera/wissrahmen?reg=r_4; 22.06.2011).

Kompetenzen

Einer verbreiteten Definition von Weinert (2001) folgend, lassen sich Kompetenzen beschreiben als „die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können." Klieme & Leutner (2006) definieren in einer für die Bildungsforschung bedeutsamen Definition „Kompetenzen als kontextspezifische kognitive Leistungsdispositionen, die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen.", betonen also die kognitiven Komponente in Abgrenzung von motivationalen und emotionalen Aspekten. Gemeinsame Merkmale beider Definitionen sind die Orientierung an prinzipiell erfassbaren Leistungen (Fähigkeiten und Fertigkeiten) sowie der Kontextbezug (Kompetenzen sind *domänenspezifisch*). In Abgrenzung zu Intelligenz erweisen sich Kompetenzen außerdem in einem deutlich stärkeren Maße als lern- und trainierbar. Köller (2009) verweist auf ein pragmatisches Verständnis von Kompetenzen in den Bildungsstandards: Kompetenz ist hier durch das „gezeigte Verhalten" in Bezug auf die Anforderungen der Standards definiert. Eine umfassende Diskussion des Kompetenzbegriffs in der Bildungsforschung nehmen Klieme, Hartig & Rauch (2008) vor.

Entwicklung der Tests, Durchführung und Ergebnisinterpretation

Die Erstellung der Testinstrumente wurde 2009 zentral durch das IQB übernommen, wobei die Projektgruppe VERA (http://www.uni-landau.de/vera/) für einige Bundesländer eine internetbasierte Ergebnisrückmeldung der Vergleichsarbeiten durchführt (vgl. Kapitel 2.4). Die Durchführung der Arbeiten und die Aufgabenbewertung erfolgen auf Basis standardisierter Manuale dezentral in den Schulen durch Lehrerinnen und Lehrer. Die Testauswertung und Ergebnisrückmeldung wird wiederum zentral durch die verantwortlichen Ministerien und Landesinstitute in Kooperation mit dem IQB (und gegebenenfalls

1 Vergleichsarbeiten sollen aufgrund der begrifflichen Ähnlichkeit in einigen Bundesländern von *Parallelarbeiten* abgegrenzt werden. Bei diesen bearbeiten die Parallelklassen einer Schule dieselben Aufgabensätze, um so einen Leistungsvergleich innerhalb einer Schule über den Klassenverband hinaus zu ermöglichen. Diese sind in der hier vorgestellten Notation nicht mit *Vergleichsarbeiten* gemeint.

2 In einigen Bundesländern erfolgen weitere Formen der Lernstandsdiagnose. Siehe hierzu Kapitel 2.4.

dem VERA-Projektteam) vorgenommen. Die Ergebnisinterpretation liegt schließlich in den Händen der Fachkonferenzen in den Schulen, wobei diese mit umfangreichen Informationen, insbesondere ausführlichen didaktischen Kommentierungen, versorgt werden, welche die Interpretation erleichtern.

Bei Vergleichsarbeiten wird in den sprachbezogenen Inhaltsbereichen eine inhaltliche Schwerpunktsetzung auf einen oder mehrere Teilleistungs- oder Kompetenzbereiche mit dem Ziel vorgenommen, die Ausprägung der getesteten Kompetenzen in zu ermitteln (Tabelle 1). Die Schwerpunktsetzung innerhalb eines Faches wechselt von Jahr zu Jahr, sodass die gesamte Breite dieses Faches im Laufe weniger Jahre über die verschiedenen Kompetenzbereiche hinweg abgebildet werden kann. Zum Einsatz kommen Aufgaben- bzw. Itemstichproben, welche repräsentativ für die jeweiligen Kompetenzerwartungen sind und zumeist einheitlich innerhalb eines Bildungsganges eingesetzt werden. Beabsichtigt ist es bei Vergleichsarbeiten nicht, eine möglichst reliable Schätzung der Kompetenzausprägung für eine einzelne Schülerin bzw. einen einzelnen Schüler zu erzielen (Individualdiagnostik; Studienbrief 1) oder aber anhand von Stichproben die Ausprägungen von Kompetenzen auf Landes- oder Bundesebene in der Breite eines Faches zu erfassen (Bildungsmonitoring; Studienbrief 3).

Abgrenzung von Individualdiagnostik / Bildungsmonitoring

	Individualdiagnostik	**Vergleichsarbeiten**	**Bildungsmonitoring**
Zielsetzung	Vorbereitung von Entscheidungen im Einzelfall	Vorbereitung pädagogischer, didaktischer und/ oder curricularer Entscheidungen auf Schul- und Unterrichtsebene (Selbstevaluation)	Vorbereitung politischer Entscheidungen auf Schulsystemebene (Fremdevaluation)
Fokus	Inhaltliche Tiefe in einem Fachgebiet: → Aufgaben bzw. Itemstichproben → Schätzung der Ausprägung einer Eigenschaft einer Schülerin/eines Schülers	Zunächst Tiefe in einem Fachgebiet, ggf. dann Breite durch Abdeckung weiterer Fachgebiete eines Faches in den Folgejahren: → Aufgaben bzw. Itemstichproben → Kompetenzerhebung der Schülerinnen und Schüler in Klassen und Jahrgangsstufen (keine Individualdiagnostik)	Fachliche Tiefe und zugleich fachliche Breite: → Aufgaben bzw. Itemstichproben und Stichprobe der Schülerinnen und Schülern (Multiple-Matrix-Stichprobe; vgl. Studienbrief 3) → Kompetenzerhebung in Bundesländern, Staaten (keine Individualdiagnostik, keine Aussagen auf Schul- und Klassenebene)

Tabelle 1: Zielsetzung von Individualdiagnostik, Vergleichsarbeiten und Bildungsmonitoring

In Kapitel 2.4 wird erläutert, wie sich die aktuelle Konzeption von Vergleichsarbeiten in den jeweiligen Unterrichtsfächern darstellt.

2.1.3 Allgemeine Zielsetzungen

allgemeine Zielsetzungen

Die Bundesländer haben in Bezug auf Vergleichsarbeiten eine unterschiedlich lange Tradition mit teilweise unterschiedlicher Gewichtung der Zielsetzungen. Übereinstimmend lassen sich aber folgende allgemeine Zielsetzungen von Vergleichsarbeiten festhalten (vgl. hierzu auch das Arbeitspapier der *Arbeitsgruppe Empirische Schulentwicklung, EMSE*, aus dem Jahr 2006):

- *Bestandsaufnahme fachlicher Kompetenzen und erreichter Lernstände, primär auf der Ebene der Klassen und Schulen*
- *didaktische und pädagogische Impulssetzung für eine datengestützte Unterrichtsentwicklung*
- *Identifikation von Förderbedarf in Lerngruppen*
- *Entwicklung und Stärkung der diagnostischen Kompetenz von Lehrkräften, insbesondere Stärkung der kriterialen Perspektive orientiert an Standards*

2.1.3.1 Bestandsaufnahme fachlicher Kompetenzen und erreichter Lernstände, primär auf der Ebene der Klassen und Schulen

Bestandsaufnahme fachlicher Kompetenzen und erreichter Lernstände

Bei Vergleichsarbeiten wird die Ausprägung von Kompetenzen ermittelt. Die Rückmeldung der Ergebnisse aus Vergleichsarbeiten umfasst nicht nur rein numerische Aussagen zum Ergebnis der Lerngruppe. Vielmehr sind die Rückmeldungen so aufbereitet, dass sie kriteriale Vergleiche (vgl. Studienbrief 1; siehe 2.2.1) im Hinblick auf das Erreichen bestimmter Leistungsanforderungen ermöglichen. Bei der Rückmeldung als Verteilung von Schülerinnen und Schülern auf *Kompetenzstufen*, werden diese Kompetenzstufen durch konkrete Aufgabenbeispiele inhaltlich so beschrieben, dass eine klare Vorstellung davon vermittelt werden kann, welchen Anforderungen die Schülerinnen und Schüler gerecht werden. Die Anforderungen der Testaufgaben sind durch die nationalen Bildungsstandards definiert und geben damit eine objektive Rückmeldung über das Erreichen der dort festgeschriebenen Anforderungen des jeweiligen Unterrichtsfaches. Vergleichsarbeiten ermöglichen den Schulen damit, im Vorfeld des Abschlusses der Primarstufe bzw. der Sekundarstufe I, eine schulinterne Standardüberprüfung.

2.1.3.2 Didaktische und pädagogische Impulssetzung für eine datengestützte Unterrichtsentwicklung

didaktische/ pädagogische Impulssetzung

Flächendeckend durchgeführte Vergleichsarbeiten haben die Funktion, pädagogische, didaktische und curriculare Entscheidungen auf der Ebene der Einzelschule anzustoßen. Im Rahmen der Ergebnisinterpretation müssen in den Schulen, zumeist innerhalb der Fachkonferenzen, die Ursachen für einen hohen oder niedrigen Ausprägungsgrad der getesteten Kompetenzen erkundet werden. Idealerweise werden in diesem Prozess Konsequenzen und Maßnahmen diskutiert, wie der Kompetenzerwerb der Schülerinnen und Schüler besser gefördert bzw. ein hohes Leistungsniveau auf lange Zeit gehalten werden kann. Differenzen in den erreichten Anforderungen zwischen den Fächern, zwischen Teilleistungsbereichen innerhalb eines Faches, zwischen den Klassen derselben Jahrgangsstufe einer Schule oder im Vergleich zu entsprechenden Referenzgruppen können als Anknüpfungspunkte für inhaltliche Schwerpunktsetzungen dienen. Mit diesen Erwartungen ausgestattet, entsprechen Vergleichsarbeiten tendenziell eher dem Konzept der Selbstevaluation (siehe Erläuterungen unten), da zwar Testverfahren zentral erstellt, diese aber dezentral in den Schulen von den Lehrkräften selbst eingesetzt, kodiert und bezüglich ihrer Ergebnisse interpretiert werden.

Selbst- vs. Fremdevaluation im Bildungswesen

Ruep und Keller (2007) beschreiben das Verhältnis von Selbst- und Fremdevaluation als „Vergleich zweier Wahrnehmungen". Bei der Selbstevaluation findet eine Bewertung pädagogischer Arbeit nicht durch externe Experten statt, sondern durch die verantwortlichen Personen vor Ort (in der Regel also die Schulleitung, die didaktischen Leitung, Koordinatorinnen und Koordinatoren, Fachkonferenzen). Die Bezeichnung Selbstevaluation beinhaltet allerdings nicht, dass der gesamte Evaluationsprozess von verantwortlichen Personen der Schulen übernommen werden muss; jedoch bleibt die Schule selbst verantwortlich für die spezifischen Ziele der Evaluation, die Verwendung und Interpretation ihrer Ergebnisse sowie die Auswahl und Umsetzung von Interventionsmaßnahmen. Die umgesetzten Maßnahmen bedürfen wiederum einer Bewertung, sodass sich häufig ein Zyklus aus Evaluations- und Interventionsschritten ergibt. Ziel der Selbstevaluation ist die Feststellung der Schul- und Unterrichtsqualität. Zur Fragestellung kann aber auch die optimale Nutzung vorhandener Ressourcen (Sach- und Personalmittel) werden. Die Ergebnisse der Selbstevaluation können außerdem zum Ausgangspunkt fremdevaluierender Maßnahmen werden oder aber diese ergänzen. Bei der *Fremdevaluation* entsteht durch die Bestandsaufnahme kritischer, externer Evaluatoren (in der Regel Mitglieder der Schulaufsicht oder anderer Schulen) ein objektives Bild des Lehrens und Lernens. Charakteristisch ist deren Bewertung von gut operationalisierten Indikatoren der Schul- und Unterrichtsqualität mit Hilfe standardisierter Verfahren der Beobachtung und Befragung. Ziele und Zeitpunkt der Evaluation werden mit der Schulleitung im Vorhinein abgeklärt. Die Schulleitung und das Kollegium werden nach Auswertung quantitativer und qualitativer Daten durch einen detaillierten Evaluationsbericht über die Ergebnisse unterrichtet und sollten im Anschluss die Möglichkeit zur Rückmeldung erhalten. Der Evaluationsbericht geht neben der Schule auch der Schulaufsicht zu, die weitere Maßnahmen der Qualitätssicherung mit der Schule diskutiert.

2.1.3.3 Identifikation von Förderbedarf in Lerngruppen

Identifikation von Förderbedarf in Schülergruppen

Kriteriale Vergleichsmaßstäbe ermöglichen die Identifikation von Schülergruppen, welche die erwarteten Standards verfehlen. Ergebnisse des Bildungsmonitorings (vgl. Studienbrief 3) haben verdeutlicht, dass die Leistungsstreuung in den Ländern der Bundesrepublik Deutschland hoch ist. Legt man die Ergebnisse der PISA-Erhebung 2006 zugrunde, so erreichen ca. 20 % der 15-jährigen Schülerinnen und Schüler in Deutschland in den Kompetenzbereichen Lesen und Mathematik nicht die Kompetenzstufe 2

und weisen damit besonders dringenden Förderbedarf auf. Es steht zu befürchten, dass diese Schülerinnen und Schüler erhebliche Schwierigkeiten haben werden, beruflichen und allgemein lebensrelevanten Anforderungen gerecht zu werden (Artelt, Stanat, Schneider & Schiefele, 2001). Diese Anteile von Schülerinnen und Schülern mit erheblichen Defiziten werden von Lehrkräften tendenziell unterschätzt (siehe Abschnitt 2.1.3.4), was möglicherweise auch damit zu tun hat, dass ihnen objektive Rückmeldungen über die Leistungsfähigkeit nicht in ausreichendem Maße zur Verfügung stehen. Alltag der meisten Lehrkräfte sind klasseninterne Vergleichsmaßstäbe, welche aber den tatsächlichen Förderbedarf von Schülerinnen und Schülern bei einer entsprechend leistungsschwachen Referenzgruppe (Klasse/Kurs) verbergen können. Vergleichsarbeiten ermöglichen hingegen, den Anteil förderungsbedürftiger Schülerinnen und Schüler vor dem Hintergrund der in den Bildungsstandards definierten Kompetenzerwartungen zu bestimmen.

2.1.3.4 Entwicklung und Stärkung der diagnostischen Kompetenz von Lehrkräften, insbesondere Stärkung der kriterialen Perspektive orientiert im Hinblick auf Standards

Entwicklung und Stärkung diagnostischer Kompetenz

Diagnostische Kompetenzen von Lehrerinnen und Lehrern werden zumeist als Fähigkeiten verstanden, Schülerinnen und Schüler hinsichtlich der Ausprägung bestimmter Merkmale treffend zu beurteilen (vgl. Schrader, 2001; Spinath, 2005). Des Weiteren lässt sich die Fähigkeit der Lehrkraft, Aufgaben hinsichtlich ihrer Schwierigkeit für die jeweilige Lerngruppe angemessen einzuschätzen, unter diagnostischer Kompetenz einordnen. Referenzrahmen ist hier die tatsächliche (empirische) Schwierigkeit der Aufgabe, in der Regel also die Lösungsquote.

Reflexion diagnostischer Überlegungen

Wie aber können Vergleichsarbeiten dazu beitragen, die diagnostischen Kompetenzen von Lehrerinnen und Lehrern zu stärken? Stellt man die treffende Beurteilung von Schülerinnen und Schüler in den Vordergrund, so erlauben die mit den Ergebnisrückmeldung verbundenen sozialen Vergleichsmöglichkeiten (vgl. Kapitel 2.2.2) eine objektive Leistungseinschätzung, unabhängig vom zuvor beschriebenen klasseninternen (bzw. schulinternen) Vergleichsmaßstab. Lehrkräften bietet sich die Möglichkeit, eigene Leistungserwartungen anhand der objektiven Ergebnisse von Referenzgruppen zu korrigieren. Mehr noch steht aber die Stärkung der *kriterialen* Perspektive der Lehrkräfte im Vordergrund (vgl. Kapitel 2.2): Von ihnen wird eine möglichst treffende Verortung des Leistungsstandes der Lerngruppe in Bezug auf die nationalen Bildungsstandards erwartet, die bei Vergleichsarbeiten durch Überprüfung des tatsächlichen Lernstandes (z.B. über die Verteilung auf Kompetenzstufen) kritisch reflektiert werden kann.

Auch in Bezug auf den zweiten Aspekt diagnostischer Kompetenzen, die angemessene Einschätzung von Aufgaben, bieten die Vergleichsarbeiten einen Referenzmaßstab: Die Aufgabenbeispiele der Vergleichsarbeiten konkretisieren die Inhalte der Bildungsstandards und ermöglichen Lehrkräften somit ein besseres Verständnis dieser abstrakt formulierten Inhalte. Grundlage für die objektive Einschätzung der Schwierigkeit ist die z.B. im Jahr 2011 erfolgte Rückmeldung von Lösungsquoten für die Aufgaben in verschiedenen Bildungsgängen. Die Analyse der Testaufgaben bietet so in Bezug auf Inhalte und Schwierigkeitsgrad Anhaltspunkte für die Ableitung von Lernaufgaben, die kritische Gegenüberstellung vorhandener Unterrichtsmaterialien und die Ausarbeitung von kompetenzorientierter Unterrichtsreihen (vgl. Kapitel 2.5.4).

2.1.4 Weiterführende Literatur

Literatur

Entwicklung und Zielsetzungen von Vergleichsarbeiten sind durch die Länder auf ihren Internetseiten dargelegt worden (siehe auch http://www.iqb.hu-berlin.de/vera?reg=r_1). Als repräsentative Auswahl der Literatur zum wissenschaftlichen Hintergrund können Helmke und Hosenfeld (2003a, b), Leutner, Fleischer, Spoden und Wirth (2007), Nachtigall und Jantowski (2004) sowie Peek (2004) empfohlen werden.

2.1.5 Verständnis und Diskussionspunkte

1. *Nennen Sie Charakteristika von Vergleichsarbeiten.*
2. *Fassen Sie allgemeine Zielsetzungen von Vergleichsarbeiten zusammen. Versuchen Sie insbesondere zu skizzieren, wie diese Zielsetzungen ineinander greifen.*
3. *Legen Sie dar, wieso diese Zielsetzungen nicht durch bestehende Instrumente erfüllt werden konnten:*
    - *Klassenarbeiten*
    - *Parallelarbeiten*
    - *Instrumente des Bildungsmonitorings (IGLU, PISA, etc.).*

2.2 Bewertungskriterien bei Vergleichsarbeiten

Die Rückmeldung des absoluten Ergebnisses einer Klasse oder Jahrgangsstufe ist für sich betrachtet häufig wenig aussagekräftig und birgt die Notwendigkeit in sich, geeignete Maßstäbe für die Interpretation der Testergebnisse heranziehen zu können. In Studienbrief 1 sind in diesem Zusammenhang kriteriale, soziale und individuelle Bezugsnormen als Interpretationsmaßstäbe eingeführt worden. Im Folgenden wird verdeutlicht, welche Informationen als Bewertungskriterien zur Interpretation von Ergebnisrückmeldungen aus Vergleichsarbeiten genutzt werden können.

In diesem Kapitel werden folgende Fragen beantwortet:

- *Inwiefern stellen die nationalen Bildungsstandards eine kriteriale Bezugsnorm für die Ergebnisse aus Vergleichsarbeiten dar?*
- *Was ist unter Fairness sozialer Vergleiche zu verstehen und wie wird sichergestellt, dass soziale Vergleiche von Testergebnissen im Kontext von Vergleichsarbeiten fair gestaltet sind?*

2.2.1 Kriteriale Bezugsnorm bei Vergleichsarbeiten

Anforderungen der Bildungsstandards als Kriterium

Kriteriale Bezugsnormorientierung beschreibt die Interpretation von Testergebnissen in Bezug auf ein inhaltlich definiertes Kriterium. Die Inhalte der Testaufgaben von Vergleichsarbeiten sind durch die nationalen Bildungsstandards (KMK, 2003a-c; 2004a-e) festgelegt, welchen die landespezifischen Curricula verpflichtet sind. In den Bildungsstandards werden fachspezifisch Schülerkompetenzen festgelegt, die bis zu einem bestimmten Zeitpunkt in der Bildungsbiographie erworben werden sollen. Vergleichsarbeiten ermöglichen die schulinterne Überprüfung dieses Kompetenzerwerbs. Die Testaufgaben der Vergleichsarbeiten werden aus fachdidaktischer, pädagogischer und lehr-lernpsychologischer Sicht bezüglich ihrer Anforderungen analysiert. Als Testaufgaben kommen dabei nur solche Aufgaben in Betracht, deren Lösungsanforderungen die in den Bildungsstandards definierten Kompetenzen voraussetzen (vgl. das folgende Beispiel). Aktuell werden in den meisten Kompetenzstufenmodellen Aufgaben mit ähnlichem, inhaltlich beschreibbarem Anforderungsprofil im Rahmen der Testkonstruktion zusammengefasst und unter Berücksichtigung ihrer Schwierigkeit als Kompetenzstufen ausgewiesen (vgl. Kapitel 2.3).

Von der KMK sind so genannte Mindeststandards formuliert worden, die von allen Schülerinnen und Schülern erreicht werden sollen (vgl. Kap. 2.4). Die Erfüllung der Mindeststandards stellt ein zentrales Kriterium zur Ergebnisinterpretation bei Vergleichsarbeiten dar.

Beispiel:

Aufgabenbeispiele

Die nachfolgend dargestellten Aufgaben (Abbildung 2, Abbildung 3) wurden in den nationalen Bildungsstandards als Beispielaufgaben[3] für den Kompetenzbereich „Leseverstehen" im Fach Englisch veröffentlicht (KMK, 2004d). An ihnen soll die Validität der Aufgaben in Bezug auf die Bildungsstandards verdeutlicht werden.

Das Aufgabenbeispiel 1 beinhaltet einen (diskontinuierlichen) Sachtext mit 211 Wörtern und enthält zehn Items des Itemfomats *„true/false"*. Die Aufgabe erfasst Kompetenzen, die noch bis zum Hauptschulabschluss im Fach Englisch erwartet werden.

Aufgabenbeispiel 3 umfasst einen fiktionalen Text mit 196 Wörtern. Sieben Items des Formats *„multiple-choice"* erfassen ebenfalls Kompetenzen, die überwiegend bis zum Hauptschulabschluss beherrscht werden sollen.

In den Bildungsstandards werden die den Aufgaben zugrunde liegenden Anforderungsmerkmale folgendermaßen definiert (relevante Stellen jeweils fett markiert):

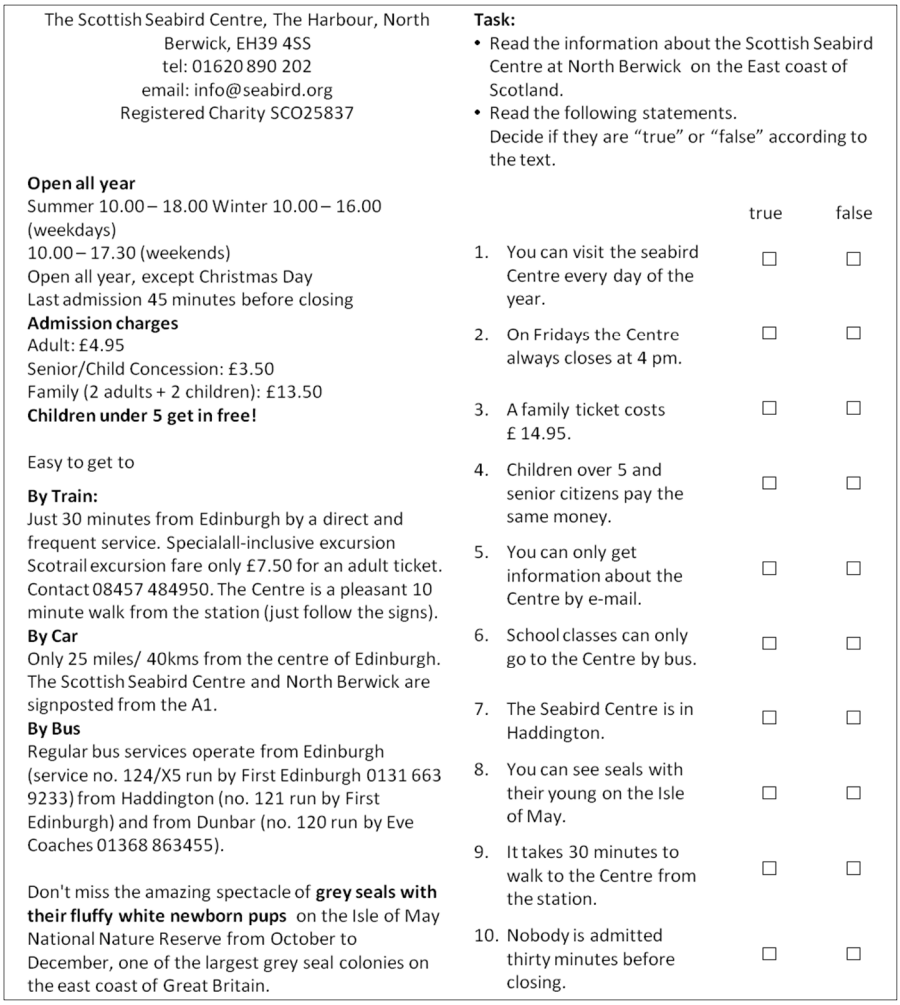
„Die Schülerinnen und Schüler können **kurze, einfache Texte lesen und verstehen**, die einen sehr **frequenten Wortschatz** und einen **gewissen Anteil international bekannter Wörter** enthalten (A 2).

Die Schülerinnen und Schüler können

- *kurze, einfache persönliche Briefe und E-Mails verstehen (A2),*
- ***konkrete, voraussagbare Informationen in einfachen Alltagstexten auffinden**, z.B. in **Anzeigen**, **Prospekten**, Speisekarten, Fahrplänen, Programmzeitschriften (A2),*
- *gebräuchliche Zeichen und Schilder an öffentlichen Orten, z.B. Wegweiser, Warnungen vor Gefahr verstehen (A2),*

3 Das Layout der Aufgaben wurde für den Zweck der besseren Darstellung angepasst.

- ***aus einfacheren schriftlichen Materialien*** *wie Briefen, Broschüren, Zeitungsartikeln (oder auch* ***dem Niveau entsprechenden fiktionalen Texten) spezifische Informationen herausfinden*** *(A2),*
- *einfache Anleitungen für Apparate verstehen, mit denen sie im Alltag zu tun haben (A2).“*



The Scottish Seabird Centre, The Harbour, North Berwick, EH39 4SS
tel: 01620 890 202
email: info@seabird.org
Registered Charity SCO25837

Open all year
Summer 10.00 – 18.00 Winter 10.00 – 16.00 (weekdays)
10.00 – 17.30 (weekends)
Open all year, except Christmas Day
Last admission 45 minutes before closing
Admission charges
Adult: £4.95
Senior/Child Concession: £3.50
Family (2 adults + 2 children): £13.50
Children under 5 get in free!

Easy to get to

By Train:
Just 30 minutes from Edinburgh by a direct and frequent service. Specialall-inclusive excursion Scotrail excursion fare only £7.50 for an adult ticket. Contact 08457 484950. The Centre is a pleasant 10 minute walk from the station (just follow the signs).
By Car
Only 25 miles/ 40kms from the centre of Edinburgh. The Scottish Seabird Centre and North Berwick are signposted from the A1.
By Bus
Regular bus services operate from Edinburgh (service no. 124/X5 run by First Edinburgh 0131 663 9233) from Haddington (no. 121 run by First Edinburgh) and from Dunbar (no. 120 run by Eve Coaches 01368 863455).

Don't miss the amazing spectacle of **grey seals with their fluffy white newborn pups** on the Isle of May National Nature Reserve from October to December, one of the largest grey seal colonies on the east coast of Great Britain.

Task:
- Read the information about the Scottish Seabird Centre at North Berwick on the East coast of Scotland.
- Read the following statements. Decide if they are “true” or “false” according to the text.

		true	false
1.	You can visit the seabird Centre every day of the year.	☐	☐
2.	On Fridays the Centre always closes at 4 pm.	☐	☐
3.	A family ticket costs £ 14.95.	☐	☐
4.	Children over 5 and senior citizens pay the same money.	☐	☐
5.	You can only get information about the Centre by e-mail.	☐	☐
6.	School classes can only go to the Centre by bus.	☐	☐
7.	The Seabird Centre is in Haddington.	☐	☐
8.	You can see seals with their young on the Isle of May.	☐	☐
9.	It takes 30 minutes to walk to the Centre from the station.	☐	☐
10.	Nobody is admitted thirty minutes before closing.	☐	☐

Abbildung 2: Aufgabenbeispiel 1 der Bildungsstandards für den Kompetenzbereich Leseverstehen im Fach Englisch in Anlehnung an KMK (2004d, S. 26 f.)

2.2.2 Soziale Bezugsnorm bei Vergleichsarbeiten

soziale Bezugsnorm

Soziale Vergleichsmaßstäbe für Lehrerinnen und Lehrer beschränken sich häufig auf klassen- oder jahrgangsstufeninterne Leistungsvergleiche und vernachlässigen schulübergreifende Vergleichsmöglichkeiten (Schrader & Helmke, 2001). Um die Leistungsfähigkeit einer Lerngruppe allerdings realistisch einschätzen zu können, ist es notwendig, ihren Lernstand mit Lerngruppen anderer Schulen vergleichen zu können. Für diesen sozialen Vergleich werden bei Vergleichsarbeiten Klassen oder Kurse ausgewählt, die sich aufgrund ihrer Ähnlichkeit in leistungsrelevanten Merkmalen als Referenzgruppe besonders eignen (vgl. Leutner, 2010). Vergleichsarbeiten sind zwar nicht vollends um die Strenge oder Milde im Urteil der Lehrkraft bereinigt (vgl. Leutner, Fleischer, Spoden & Wirth, 2007; vgl. Kapitel 2.3), Unterschiede in der persönlichen Urteilsstrenge zwischen Lehrerinnen und Lehrern werden aber aufgrund desselben Aufgabenmateriales in allen Referenzklassen, der überwiegenden Verwendung geschlossener Aufgabenformate und einer standardisierten Durchführungs- und Auswertungsanleitung minimiert. Damit soziale Vergleiche aber objektive Aussagen über die Qualität der Beschulung zulassen, ist es zudem notwendig, Schulen faire Referenzwerte bereitzustellen.

2.2.2.1 Faire Vergleiche

faire Vergleiche

Schulische Lerngelegenheiten stellen notwendige Voraussetzungen für den Aufbau von Schülerkompetenzen dar und die Qualität dieser Lerngelegenheiten hat einen entscheidenden Einfluss auf die Ausprä-

gung der erworbenen Kompetenzen (Köller & Baumert, 2002; Helmke & Schrader, 2001). Darüber hinaus besitzen jedoch auch Merkmale der Schülerinnen und Schüler, die sich dem Einflussbereich von Schule und Unterricht entziehen, Einfluss auf den Aufbau von Kompetenzen. Darunter sind zum einen individuelle kognitive Eingangsvoraussetzungen, insbesondere Intelligenz, zu fassen (Helmke & Schrader, 2001; Süß, 2001). Doch auch sozioökonomische Merkmale der Schülerschaft, z.B. der elterliche Bildungsgrad und die Familiensprache, besitzen nachweislich einen Einfluss auf das schulische Leistungsniveau (Ramm, Prenzel, Heidemeier & Walter, 2004; Nachtigall, Kröhne, Enders & Steyer, 2008). Ergebnisse der PISA-2000-Studie haben verdeutlicht, dass die Unterschiede in der Zusammensetzung der Schülerschaft in Bezug auf Kontextmerkmale sogar innerhalb derselben Schulform beträchtlich sein können (vgl. Baumert et al., 2003). Soziale Vergleiche einer Klasse oder Jahrgangsstufe mit Vergleichspopulationen sollten dahingehend fair gestaltet sein, als dass leistungsrelevante Kontextmerkmale der Schülerschaft beim sozialen Vergleich Berücksichtigung finden müssen. Andernfalls bestünde die Gefahr, dass an Schulen mit ungünstigen Eingangsvoraussetzungen die Effektivität der schulischen Arbeit unterschätzt, an Schulen mit günstigen Eingangsvoraussetzungen überschätzt wird (Peek & Dobbelstein, 2006). Das Ziel der Berücksichtigung von Kontext- oder Hintergrundinformationen ist daher, die tatsächliche Qualität der Beschulung von Leistungseinflüssen zu trennen, die aufgrund von Merkmalen der sozialen Herkunft oder individueller kognitiver Fähigkeiten zustande kommen.

Ingrid McFarlane
Zoo Keeper

When I left school at eighteen, I got a job at a zoo as a student keeper. Now, five years later, things have changed – I have passed my exams and I am a full animal keeper.

The money is not good. I only get £9,000 a year. You have to be outside in rain and snow, which is hard work, and you get very dirty. But this doesn't matter to me because animals are the most important thing in my life!

There are a hundred monkeys and fifty deer in my part of the zoo and I give them their food and clean their houses. I also need to watch them carefully to be sure that they are all well. In fact, rhinos are my favourite animals and so last year I went to Africa with a colleague for a month to study them.

The zoo is open every day and I work five different days each week. I live in a small flat twenty minutes away and I get up at ten to seven and start work at eight. The first thing I do when I get home at quarter past five is have a shower!

Task:

- Read the article about Ingrid McFarlane and then answer the
- For questions 1 – 7 mark A, B or C on your answer sheet.

1 Ingrid would like to
A take some exams.
B earn more money.
C change her job.

2 How does Ingrid feel about working in bad weather?
A She hates getting dirty.
B She doesn't mind it.
C She likes the snow.

3 If Ingrid doesn't check the monkeys
A they may become ill.
B they may get hungry.
C they may run away.

4 The animals Ingrid likes best are the
A monkeys.
B deer.
C rhinos.

5 Ingrid travelled to Africa
A to have a month's holiday.
B to visit a colleague there.
C to learn more about some animals.

6 The zoo is open
A only five days a week.
B seven days a week.
C on different days every week.

7 Ingrid arrives at her flat in the evening at
A five fifteen.
B twenty past five.
C ten to seven.

Abbildung 3: Aufgabenbeispiel 3 der Bildungsstandards für den Kompetenzbereich Leseverstehen im Fach Englisch in Anlehnung an KMK (2004d, S. 29 f.)

Fiktives Beispiel: Die Bedeutung unterschiedlicher Eingangsvoraussetzungen der Schülerschaft

Seit langer Zeit ist bekannt, dass im Ruhrgebiet ein soziales Nord-Süd-Gefälle besteht. Der Norden des Ruhrgebietes beinhaltet tendenziell strukturschwächere, der Süden strukturstärkere Stadtteile. Diese Differenz der Sozialstruktur spiegelt sich auch im sozioökonomischen Hintergrund der Schülerschaft zweier fiktiver Schulen wider, der Nordschule und der Südschule (Tabelle 2).

Betrachtet man die reinen numerischen Ergebnisse, so erzielt die Südschule deutlich bessere Leistungen als die Nordschule (Spalte 3). Um dem sozialen Hintergrund der Schülerschaft bei der Bewertung der schulischen Arbeit gerecht zu werden, wurden von allen Schülerinnen und Schülern zwei Sozialindikatoren erfragt: Die Anzahl von Einfamilienhäusern im Umkreis von 500 Metern um die eigene Wohnung und die Anzahl von Büchern im Haushalt der Familie (Spalte 7).

Es konnte statistisch erfasst werden, dass die Testleistung aller Schülerinnen und Schüler bei einer Vergleichsarbeit pro Einfamilienhaus im Wohnungsumfeld um (durchschnittlich) 0,5 Testpunkte und pro Buch im Haushalt um 0,02 Testpunkte zunimmt. Um einen fairen Vergleich zu ermöglichen, wurden die tatsächlichen Testergebnisse daher in zwei Schritten am jeweiligen Mittelwert der Anzahl von Einfamilienhäusern im Umkreis der eigenen Wohnung und der Anzahl von Büchern im Haushalt korrigiert. Im ersten Korrekturschritt wurde das Testergebnis für jedes zusätzliche Einfamilienhaus im Umkreis der eigenen Wohnung größer dem Mittelwert 4 um jeweils 0,5 Testpunkte reduziert und für jedes Einfamilienhaus kleiner 4 um die gleiche Punktzahl erhöht (Spalte 4). Es ist anhand der Mittelwerte pro Schule ersichtlich, dass sich die korrigierten Testleistungen der beiden Schulen bereits angenähert haben. Im zweiten Schritt wurde nun der Faktor „Bücher im Haushalt“ als Indikator für kulturelles Kapital berücksichtigt. Dazu wurden die Testleistungen um die Differenz der sich im Haushalt befindlichen Bücher gegenüber dem Mittelwert von 90 Büchern korrigiert (Spalte 5). Nach diesem Korrekturschritt schneidet die Nordschule nun sogar besser ab als die Südschule. Berücksichtigt man also Unterschiede der Schülerschaft, so kann die Qualität der Beschulung in einem ganz anderen Licht erscheinen.

	Schüler	Test-leistung	Korrektur 1	Korrektur 2	Anzahl der Einfamili-enhäuser im Wohn-umfeld	Anzahl der Bücher im Haushalt
Nordschule	Schüler 1	6	6	7,2	4	30
	Schülerin 2	4	5	6,5	2	15
	Schüler 3	10	12	12,8	0	50
	Schülerin 4	12	13	14,3	2	25
Südschule	Schüler 5	14	12	11,2	8	130
	Schülerin 6	16	15	13	6	190
	Schüler 7	10	9	7,2	6	180
	Schülerin 8	8	8	7,8	4	100
Mittelwerte	gesamt	10	10	10	4	90
	Nordschule	8	9	10,2	2	30
	Südschule	12	11	9,8	6	150

Tabelle 2: Fiktives Beispiel für die Adjustierung von Testleistungen aufgrund des sozioökonomischen Hintergrundes der Schülerschaft

Die statistische Korrektur von Leistungen für den sozialen Vergleich impliziert natürlich nicht, dass die Schülerinnen und Schüler der Nordschule in Bezug auf das Lehrziel „mehr können“, als sie im Test gezeigt haben. Die durchschnittliche Anzahl gelöster Items liegt in der Nordschule bei 8, und die Schülerinnen und Schüler der Nordschule haben tatsächlich im Sinne eines kriterialen Vergleichs „Aufholbedarf“ gegenüber der Südschule. Dieser „Aufholbedarf“ ist aber nicht das Resultat ungünstiger Beschulung, sondern ungünstiger Eingangsvoraussetzungen der Schülerschaft an der Nordschule.

faire Vergleiche – Methoden

Um „faire Vergleiche“ unter Berücksichtigung leistungsrelevanter Kontextmerkmale der Schülerschaft zu realisieren, lassen sich zwei grundlegende Konzepte[4] unterscheiden (vgl. für eine ausführlichere Darstellung Nachtigall, Kröhne, Enders & Steyer, 2008):

4 Nachtigall, Kröhne, Enders und Steyer (2008) unterscheiden neben dem Vergleich mit dem (ungewichteten) Populationsmittelwert drei Verfahren: ‚comparison to similar existing classes or schools‘, ‚comparison to expected values‘ und ‚comparison based on propensity scores‘. Aus didaktischen Gründen sind an dieser Stelle die letzten beiden Verfahren zusammengefasst worden.

1. Der Vergleich mit Klassen (bzw. Schulen), deren Schülerschaft möglichst große Ähnlichkeit in Bezug auf sozioökonomische Merkmale besitzt. Schulen werden dazu sogenannten Belastungs-, Kontext- oder Standortgruppen zugeordnet.
2. Der Vergleich mit statistisch korrigierten Leistungswerten, die tatsächliche Schulleistungen in Beziehung zu Schulleistungen setzen, die aufgrund des sozioökonomischen Hintergrundes der Schülerschaft zu erwarten wären. Dazu werden die Testergebnisse statistisch „aufgewertet", wenn die Schülerzusammensetzung der Klasse ungünstiger als in Referenzklassen ist. Die Testergebnisse werden hingegen „abgewertet", wenn die Zusammensetzung der Schülerschaft gegenüber Referenzklassen als günstiger zu bezeichnen ist.

Das erste vorgestellte Konzept bietet Lehrerinnen und Lehrern insbesondere den Vorteil höherer Transparenz und Nachvollziehbarkeit des sozialen Vergleichs, beinhaltet aber auch die Herausforderung, geeignete Vergleichsklassen oder -schulen zu identifizieren. Bei Anwendung des zweiten Konzepts lassen sich genauere Ergebnisse ermitteln; deren Zustandekommen ist für die Adressaten der Rückmeldung aber häufig wenig durchschaubar. Zu beiden Konzepten ein kurzes Beispiel:

Beispiele:

faire Vergleiche – Beispiele

Im Rahmen der Ergebnisrückmeldung der Lernstandserhebungen in der achten Jahrgangsstufe in Nordrhein-Westfalen (lernstand 8) erhalten Schulen als Referenzwerte für den sozialen Vergleich die Testergebnisse der jeweiligen Schulform sowie die Testergebnisse so genannter Standorttypen. Jede Schule ist auf Basis von Daten der amtlichen Statistik[5] einem von insgesamt fünf Standorttypen zugeordnet worden. Als relevante Daten wurden berücksichtigt:

- Anteil der Migrantinnen und Migranten innerhalb der Schülerschaft
- Anteil der Arbeitslosen und der SGB II-Empfängerinnen und Empfänger unter 18 Jahren im Schulumfeld

Schulen des Standorttyps 1 sind durch einen niedrigen Anteil von Empfängerinnen und Empfängern staatlicher SozialhiIfeleistungen, von Arbeitslosen und von Menschen mit Migrationshintergrund charakterisiert; Schulen des Standorttyps 5 befinden sich hingegen eher in einer Lage mit schwierigen sozialen und kulturellen Rahmenbedingungen (Isaac, 2011). Sozialräumliche Veränderungen, wie etwa in der Bevölkerungsstruktur, werden durch regelmäßige Aktualisierung der amtlichen Daten berücksichtigt.

Bei den Kompetenztests (http://www.kompetenztest.de) in den sechsten und achten Klassen in Thüringen werden die durchschnittlichen Testergebnisse von Klassen mit adjustierten Landesmittelwerten verglichen. Bei den Schülerinnen und Schülern werden die folgenden Merkmale zum Zweck der Adjustierung der individuellen Testleistung erhoben (Nachtigall, Kröhne & Müller, 2005):

- Geschlecht
- Muttersprache
- Wiederholung einer Klassenstufe
- Besondere Lernschwierigkeiten bzw. sonderpädagogischer Förderbedarf
- Anzahl der Bücher im Elternhaus

Auf Basis dieser Werte wird für jede Schülerin und jeden Schüler ein Erwartungswert[6] berechnet; der durchschnittliche Erwartungswert in einer Klasse wird als Indikator für die zu erwartende Klassenleistung („korrigierter Landesmittelwert") herangezogen. Lehrerinnen und Lehrer in Thüringen erhalten dementsprechend einen individuellen, auf die besondere Zusammensetzung ihrer Klasse abgestimmten Erwartungswert der Leistungsstärke ihrer Lerngruppe.

2.2.3 Individuelle Bezugsnorm bei Vergleichsarbeiten

individuelle Bezugsnorm; regelmäßige Lernstands-diagnose

Auf der Ebene der Einzelschule lassen sich die Ergebnisse der Vergleichsarbeiten für die Evaluation schulischer Entwicklungsprozesse nutzen, wenn zusätzliche diagnostische Informationen berücksichtigt werden. Materialien zur regelmäßigen Lernstandsdiagnose können an von Lehrkräften selbst gewählten Zeitpunkten für die Überprüfung der Leistungsentwicklungen und Kompetenzzuwächse eingesetzt werden (vgl. hierzu Kapitel 2.5).

5 Den zugrundeliegenden Kernel-Density-Ansatz beschreibt Schräpler (2009).
6 Die Erwartungswerte werden über die Zellenmittelwerte einer mehrfaktoriellen Varianzanalyse geschätzt, siehe Nachtigall, Kröhne, Enders & Steyer (2008).

2.2.4 Weiterführende Literatur

Literatur

Die Bildungsstandards als zentrale kriteriale Bezugsnorm von Vergleichsarbeiten können über die Webseiten der KMK als Online-Dokumente bezogen werden (siehe Kapitel 2.4). Das Lehrbuch von Klauer (1987) gilt als Klassiker für die Darstellung kriteriumsorientierten Testens in der Schule. Faire Schulvergleiche thematisiert Arnold (1999). Die Bedeutung von Hintergrundmerkmalen für die Entstehung von Leistungsheterogenität und Möglichkeiten fairer Vergleiche werden von Nachtigall, Kröhne, Enders & Steyer (2008) dargestellt.

2.2.5 Verständnis und Diskussionspunkte

1. *Wie ist die Passung zwischen den Aufgaben der Vergleichsarbeiten und den Lehrplänen zu bewerten? Inwiefern könnte eine grobe, indirekte Anbindung vorliegen?*
2. *Bei Parallelarbeiten werden Aufgaben einer schriftlichen Leistungsüberprüfung von den Fachlehrerinnen und Fachlehrern einer Jahrgangsstufe gemeinsam entwickelt und in Parallelklassen eingesetzt. Die Auswertung der Aufgabenlösungen erfolgt wiederum in Absprache zwischen den Fachlehrerinnen und Fachlehrern. Diskutieren Sie bitte die Bedeutung von Parallelarbeiten vor dem Hintergrund der sozialen Bezugsnorm. Wie können sich Vergleichsarbeiten und Parallelarbeiten gegenseitig ergänzen?*
3. *Denken Sie sich bitte in folgendes Szenario: Um eine kostenintensive Befragung von Schülerinnen und Schülern zu vermeiden, sollen die relativ leicht zugänglichen Daten der Schulstatistik herangezogen werden, um eine Adjustierung von Schulleistungsdaten (z.B. aus Vergleichsarbeiten) vornehmen zu können und den Schulen damit faire Vergleiche zu ermöglichen. Welchen Kriterien sollten die Daten der Schulstatistik genügen? Welche Schwierigkeiten können entstehen, wenn solche Daten verwendet werden?*

2.3 Testkonstruktion

Welche Bedeutung haben Kompetenzstufen und was sagen sie aus? Wie werden Schülerinnen und Schüler bei der Ergebnisrückmeldung von Vergleichsarbeiten diesen Kompetenzstufen zugeordnet? Woher weiß man, dass die Schülerinnen und Schüler die jeweiligen Aufgaben dieser Stufen mit einiger Sicherheit lösen können? Im folgenden Kapitel wird erläutert, wie aufeinander aufbauende Leistungsanforderungen durch Kompetenzstufen beschrieben und mit Hilfe sorgfältig ausgewählter Aufgaben erfassbar gemacht werden. Bei der Aufgabenerstellung und -auswahl steht die Passung in Bezug auf die Anforderungen der Bildungsstandards im Vordergrund. Außerdem wird erläutert, wie man auf Basis des Testmodells nach Rasch (Rasch, 1960) sicherstellt, dass Schülerinnen und Schüler den Anforderungen ihrer Kompetenzstufe mit einiger Sicherheit gewachsen sind. Die aus Studienbrief 1 bekannten Konzepte der Testgüte werden an verschiedenen Stellen aufgegriffen.

In diesem Kapitel werden folgende Fragen beantwortet:

- *Wie werden Leistungsanforderungen durch Kompetenzstufen beschrieben?*
- *Wie werden Schülerinnen und Schüler den Kompetenzstufen zugeordnet? Welchen Beitrag leistet das Rasch-Modell zur Konstruktion von Kompetenzskalen?*
- *Nach welchen Kriterien werden die Testaufgaben für Vergleichsarbeiten ausgewählt und wie wird sichergestellt, dass diese Kriterien eingehalten werden?*

2.3.1 Die Beschreibung von Leistungsanforderungen durch aufeinander aufbauende Kompetenzstufen

Prüfung der Testgüte

Die in den Bildungsstandards festgeschriebenen Kompetenzerwartungen werden durch eigens für diesen Zweck entwickelte und bzgl. ihrer Testgüte (vgl. Kapitel 1.3) sorgfältig überprüfte Aufgaben erfasst. Bei der Ergebnisrückmeldung werden Aufgaben, die ähnlich komplexe Anforderungen beinhalten, entsprechend ihrer Schwierigkeit gruppiert und als Kompetenzstufen inhaltlich beschrieben. Die Kompetenzstufen bauen in ihren Anforderungen aufeinander auf, d.h. Schülerinnen und Schüler, die höhere Kompetenzstufen erreicht haben, beherrschen auch die Inhalte niedrigerer Stufen sicher und lösen die Aufgaben entsprechender Kompetenzanforderungen mit hoher Wahrscheinlichkeit. Beispielhaft sei dies im Folgenden an der Beschreibung eines Kompetenzstufenmodells aus VERA (Vergleichsarbeiten in der dritten Jahrgangsstufe), Kompetenzbereich „Sachrechnen und Größen“, verdeutlicht (Helmke & Hosenfeld, 2004). Für diesen Kompetenzbereich der Primarstufen-Mathematik waren zu diesem Zeitpunkt noch drei Stufen voneinander abgegrenzt worden, die eine zunehmende Komplexität der Anforderungen (Progression) erkennen lassen.

Beispiel: Erläuterungen zu den Kompetenzstufen Mathematik: „Sachrechnen und Größen“ bei VERA 2004

Stufe 1: Elementare Kenntnisse

- Die Anwendung von Addition (auch wiederholte Additionen) und Subtraktion in authentischen Aufgaben gelingt bei Aufgaben mit Auswahl aus vorgegebenen Lösungen.
- Grundlegende Kenntnisse von vertrauten Maßeinheiten (Längen-, Zeit-, Gewichts- und Geldeinheiten).
- Offensichtlich unlösbare Aufgaben werden erkannt.

Stufe 2: Entwickelte Fähigkeiten im Umgang und Rechnen mit Größen

- Im Umgang mit vertrauten Maßeinheiten (Längen-, Zeit-, Gewichts- und Geldeinheiten) können Aufgaben bis in den Tausender-Zahlenraum gelöst werden.
- Lösungen von authentischen Aufgaben, die Umrechnungen von Maßeinheiten erfordern, gelingen.
- Rundungen und Schätzungen gelingen bei Aufgaben mit vorgegebenen Lösungen.
- Verknüpfungen von Operationen werden bewältigt.
- Der Umgang mit elementaren Brüchen gelingt.
- Aufgaben mit mehreren zu verarbeitenden Größen werden gemeistert.

Stufe 3: Eigenständige Problemlösungen

- Unlösbare Aufgaben, die eine mentale Vorstellung des geschilderten Szenarios erfordern, werden erkannt.
- Bei Aufgaben ohne vorgegebene Fragestellung kann eigenständig eine Aufgabe formuliert und bearbeitet werden.
- Funktionale Beziehungen zwischen Maßen können eigenständig hergestellt und verglichen werden.
- Die mathematische Modellierung problemhaltiger Sachsituationen gelingt.
- Aufgaben, die mehrere Teilschritte umfassen, werden beherrscht.

Während Schülerinnen und Schüler auf Stufe 1 beispielsweise lediglich „Strichrechnung“ sicher anwenden können und grundlegende Kenntnisse über Größen und Maßeinheiten besitzen, können auf Stufe 2 bereits Brüche bearbeitet werden, werden Maßeinheiten in einem breiten Zahlenraum beherrscht und mehrere Operationen und Größen verknüpft. Schülerinnen und Schüler auf Stufe 3 sind darüber hinaus in der Lage, Größen und Maße zueinander in Beziehung zu setzen, eigenständige Modellierungen vorzunehmen und mehrschrittige Aufgaben zu lösen.

2.3.2 Definition von Kompetenzstufen über Lösungswahrscheinlichkeiten: das Rasch-Modell

Rasch-Modell

Für die Definition von Kompetenzstufen ist es wichtig sicherzustellen, dass Schülerinnen und Schüler der richtigen Kompetenzstufe zugeordnet werden, also die jeweiligen Anforderungen bewältigen und entsprechende Aufgaben hinreichend sicher lösen. Im sogenannten Rasch-Modell (Rasch, 1960) aus der probabilistischen Testtheorie (Rost, 2004) wird die Wahrscheinlichkeit einer korrekt gelösten Aufgabe als eine (logistische) Funktion der Schülerkompetenz (θ) und der Schwierigkeit der Aufgabe (σ) geschätzt und kann mit Hilfe einer so genannten Item-charakteristischen Kurve (ICC) veranschaulicht werden. Beispielhaft werden in Abbildung 4 die ICCs für zwei Testaufgaben dargestellt. In der Abbildung sind die Aufgabenschwierigkeit bzw. Schülerkompetenz (beide werden im Rasch-Modell auf einer Skala abgebildet) gegen die Lösungswahrscheinlichkeit aufgetragen. Es ist ersichtlich, dass die ICCs für beide Aufgaben dieselbe Steigung besitzen. Aufgabe 2 ist also – unabhängig von der Schülerkompetenz, die hier beispielhaft für einen Schüler 1 mit geringer und einen Schüler 2 mit hoher Kompetenzausprägung veranschaulicht wurde – stets die schwierigere der beiden Testaufgaben. Als Aufgabenschwierigkeit im Rasch-Modell wird bei Vergleichsarbeiten zumeist derjenige Punkt der ICC gewählt, bei dem die Lösungswahrscheinlichkeit für die Auf-gabe bei 62,5 % liegt. Um die Anforderungen der Aufgabe 1 sicher zu lösen, müsste ein Schüler auf der Kompetenzskala also (mindestens) θ=-2,2 erreichen, für die schwierigere Aufgabe 2 mindestens θ=0,7. Eine bemerkenswerte Eigenschaft des Rasch-Modells ist die Tatsache, dass für die Schätzung der Kompetenzausprägung der Schülerinnen und Schüler θ lediglich die Summe gelöster Items unabhängig von deren Schwierigkeit relevant ist.

Die Kompetenzstufen werden – einem Konzept aus dem PISA-Kontext folgend – so gebildet, dass Schülerinnen und Schüler mit der niedrigsten Kompetenz auf jeder Stufe (niedrigste Schülerkompetenz innerhalb dieser Stufe) eine Lösungswahrscheinlichkeit von mindestens 50 % für alle Aufgaben der Stufe und eine Lösungswahrscheinlichkeit von 62,5 % für die leichteste Aufgabe der Stufe besitzen (in Abbildung 5 in blauer Farbe dargestellt). Durchschnittlich kompetente Schülerinnen und Schüler dieser Stufe (Mittelwert der Schülerkompetenz innerhalb der Kompetenzstufe) lösen eine durchschnittlich schwierige Aufgabe der gleichen Stufe mit einer Wahrscheinlichkeit von 62,5 % (in Abbildung 5 in roter Farbe dargestellt). Es wird so sichergestellt, dass Aufgaben einer Kompetenzstufe von den Schülerinnen und Schülern dieser Stufe mit einiger Sicherheit gemeistert werden.

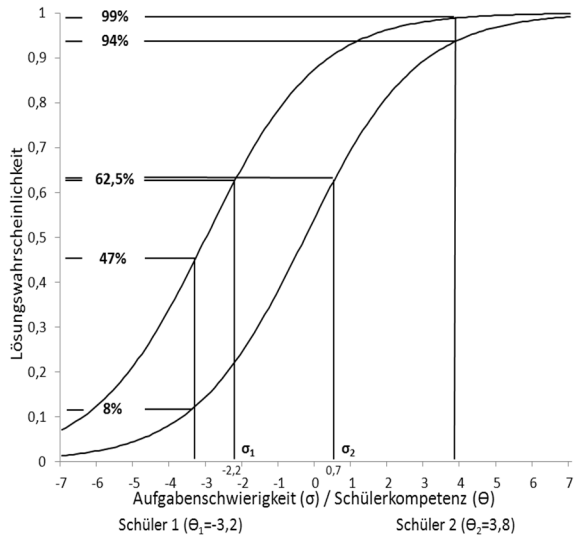


Abbildung 4: Itemcharakteristische Kurven (ICC) zweier Testaufgaben nach dem dichotomen Rasch-Modell – Zusammenhang von Aufgabenschwierigkeit (σ), Schülerkompetenz (θ) und Lösungswahrscheinlichkeit

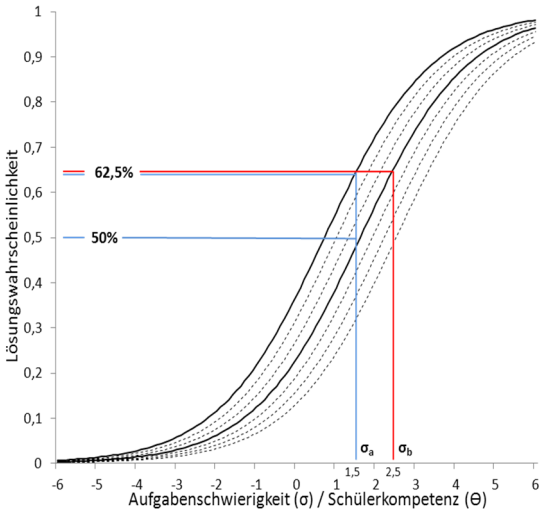


Abbildung 5: Lösungswahrscheinlichkeiten für Aufgaben einer Kompetenzstufe durch Schülerinnen und Schüler mit der niedrigsten Kompetenz innerhalb der Stufe (in blauer Farbe) und Schülerinnen und Schüler durchschnittlicher Kompetenz innerhalb der Stufe (in roter Farbe). Die Aufgabe mit der niedrigsten

2.3.3 Aufgabenauswahl bei Vergleichsarbeiten: Orientierung an Bildungsstandards und hohe Testgüte

Aufgabenerstellung/ Aufgabenauswahl

Die Qualität eines Kompetenzstufenmodells ist auch von der Auswahl geeigneter Testaufgaben abhängig. Die Aufgaben der Vergleichsarbeiten sollen die Bildungsstandards widerspiegeln, dürfen die jeweilige fachdidaktische Tradition nicht konterkarieren und müssen über eine entsprechende Testgüte verfügen (vgl. Burkard & Peek, 2004; Dobbelstein & Peek, 2007; Fleischer, Wirth & Leutner, 2007). Hinzu kommt, dass die intendierten sozialen Vergleiche (Kapitel 2.2) nur sinnvoll sind, wenn Lehrkräfte gleiche Aufgabenlösungen gleich auswerten, die Aufgaben also objektiv (auswertbar) sind. Um diesen Ansprüchen gerecht werden zu können, werden alle Testaufgaben unter realen Bedingungen mehrschrittig erprobt: Der erste Schritt beinhaltet eine Präpilotierung, bei der die Schwierigkeit der Aufgaben und die Güte der Auswertungsanleitungen grob abgeschätzt werden. Darauf folgt eine Pilotierungsphase, in der eine ausreichend große Stichprobe von Schülerinnen und Schülern die Testaufgaben bearbeitet. Die Testgüte der Aufgaben wird folgendermaßen sichergestellt:

1. Die Objektivität der Aufgabenbewertungen wird durch eine hohe Anzahl von Aufgaben geschlossener Formate erhöht. Zudem kann mit Hilfe statistischer Indizes (Wirtz & Caspar, 2002) die Übereinstimmung von mindestens zwei Lehrkräften bei der Beurteilung von Aufgabenlösungen derselben Person erfasst werden.
2. Wie auch in der klassischen Testtheorie (siehe Studienbrief 1) wird die Reliabilität der aus den Aufgaben zusammengesetzten Skala geprüft. Tests der bei Vergleichsarbeiten überwiegend eingesetzten probabilistischen Testtheorie (siehe oben) nutzen dazu Methoden analog zum in Studienbrief 1 vorgestellten Koeffizienten Cronbachs Alpha. Die Reliabilität der Vergleichsarbeiten ist hinreichend hoch, um zuverlässige diagnostische Aussagen auf der Ebene von Schulklassen zu treffen. Individualdiagnostische Aussagen (vgl. Studienbrief 1) verbieten sich allerdings weitestgehend (Leutner, Fleischer, Spoden & Wirth, 2007). Um Kompetenzen einzelner Schülerinnen

und Schüler für diese Zielsetzung genau genug zu erfassen, wäre eine höhere Item-Anzahl notwendig. Dies sollte bei der Ergebnisrückmeldung an die Schülerinnen und Schüler beachtet werden.

3. Die Validität kann – wenn überhaupt – nur grob erfasst werden. Vereinzelt kann aber geprüft werden, ob sich in den Ergebnissen bekannte Befunde (z.B. aus dem Bildungsmonitoring, vgl. Studienbrief 3) widerspiegeln, etwa Schulform- oder Geschlechtsunterschiede sowie Zusammenhänge mit sozioökonomischen Variablen.

Die Aufgabensätze der Vergleichsarbeiten stellen schließlich einen Kompromiss zwischen fachdidaktischen und testtheoretischen Anforderungen dar, d.h. inhaltliche Qualität und Testgüte müssen gleichermaßen gewährleistet sein.

Das deutsche Schulsystem weist eine breite Leistungsstreuung zwischen Schulen (insbesondere zwischen Schulen unterschiedlicher Schulformen), aber eine verhältnismäßig geringe Streuung innerhalb der Einzelschule auf (Baumert, Trautwein & Artelt, 2003). Diesem Ergebnis wird bei Vergleichsarbeiten einerseits durch die Verwendung unterschiedlich schwieriger Testheftversionen, andererseits durch ein breites Schwierigkeitsspektrum der Testaufgaben innerhalb der Hefte Rechnung getragen. Es wird so sichergestellt, dass die Testaufgaben Schülergruppen nicht systematisch über- oder unterfordern. Innerhalb einer Lerngruppe werden identische Aufgaben verwendet, sodass gut nachvollziehbare vergleichende Aussagen über die getesteten Kompetenzen getroffen werden können. Ein typisches Missverständnis bei der Interpretation von Ergebnissen aus Vergleichsarbeiten verbirgt sich in der Annahme, dass Leistungen über verschiedene Schulformen hinweg nicht vergleichbar seien. Dies trifft auf Vergleichsarbeiten in der Regel nicht zu[7]. Bei der Testauswertung auf Basis des zuvor dargestellten Rasch-Modells sind die Ergebnisse von Klassen unterschiedlicher Schulformen auch dann noch vergleichbar, wenn in den Schulformen unterschiedliche Testhefte mit zum Teil anderen Aufgaben bearbeitet wurden.

2.3.4 Weiterführende Literatur

Literatur

Das Vorgehen bei der Aufgabenentwicklung und Aufgabenselektion wird ausführlich von Leutner, Fleischer, Spoden & Wirth (2007) beschrieben. Das Rasch-Modell erläutern Bond & Fox (2001) weitestgehend „formelfrei“ und Rost (2004) auf Schulmathematik-Niveau. Die Definition von Kompetenzstufen auf der Rasch-Skala wird von Fleischer, Wirth & Leutner (2007) dargelegt, kann darüber hinaus aber auch den PISA-Bänden (z.B. Artelt, Stanat, Schneider & Schiefele, 2001) oder ebenfalls Leutner, Fleischer, Spoden & Wirth (2007) entnommen werden.

2.3.5 Verständnis und Diskussionspunkte

1. *Fassen Sie bitte zusammen, nach welchen Kriterien die Qualität der Aufgaben in Vergleichsarbeiten geprüft wird.*
2. *Beschreiben Sie bitte, weshalb Testaufgaben bei Vergleichsarbeiten objektiv auswertbar sein sollen. Überlegen Sie, wie Sie selbst – ohne Aneignung weiteren Wissens – die Auswertungsobjektivität von Aufgaben (z.B. bei Parallelarbeiten) überprüfen könnten.*
3. *Diskutieren Sie bitte, welche Informationen als Hinweise auf die Validität eines für Vergleichsarbeiten konstruierten Tests herangezogen werden könnten.*

2.4 Inhaltlicher Anwendungsbereich/Phänomenbereich[8]

VERA 3/ VERA 8

Vergleichsarbeiten werden in der dritten Jahrgangsstufe (VERA 3) in den Fächern Deutsch und Mathematik geschrieben. In der achten Jahrgangsstufe finden Vergleichsarbeiten (VERA 8) in den Fächer Deutsch, Mathematik und der ersten Fremdsprache (Englisch oder Französisch) statt. Die Webseite des IQB (http://www.iqb.hu-berlin.de/vera?lang=en&reg=r_1) weist Vergleichsarbeiten in allen Bundesländern der BRD aus. Aktuelle Informationen zu Vergleichsarbeiten in jedem Bundesland sind den beim IQB aufgeführten Webseiten der Länder zu entnehmen.

Dieses Kapitel umfasst Hintergrundinformationen zu den Kompetenzstufenmodellen in zwei bzw. drei Fächern der Primar- und Sekundarstufe[9]. Die Leserinnen und Leser sind eingeladen, die jeweils für sie relevanten Unterkapitel auszuwählen:

7 Voraussetzung ist allerdings, dass ein gemeinsames Kompetenzstufenmodell zugrunde liegt. Derzeit stellt das Kompetenzstufenmodell für den mittleren Schulabschluss (MSA) den Referenzrahmen dar.

8 Stand: 22.06.2011; vgl. http://www.iqb.hu-berlin.de/bista?reg=r_4.

9 Die nachfolgende Darstellung der Kompetenzstufenmodelle bei VERA 3 und VERA 8 spart notwendigerweise Details aus. Die Beschreibung der Vergleichsarbeiten in der ersten Fremdsprache wurde auf das Fach Englisch beschränkt, in den Fächern Deutsch und Englisch wurde jeweils ein Kompetenzbereich ins Zentrum gerückt.

In diesem Kapitel werden folgende Fragen beantwortet:

- *Welche Inhalte werden bei den Vergleichsarbeiten in der dritten Jahrgangsstufe (VERA 3) in den Fächern Deutsch und Mathematik thematisiert?*
- *Wie werden diese Inhalte bei VERA 3 operationalisiert und welche Aufgabenformate kommen zum Einsatz?*
- *Welche Inhalte werden bei den Vergleichsarbeiten in der achten Jahrgangsstufe (VERA 8) in den Fächern Deutsch, Mathematik und der ersten Fremdsprache thematisiert?*
- *Wie werden diese Inhalte bei VERA 8 operationalisiert und welche Aufgabenformate kommen zum Einsatz?*

2.4.1 Vergleichsarbeiten in der dritten Jahrgangsstufe (VERA 3)

VERA 3

Vergleichsarbeiten sind seit 2004 in nahezu allen Bundesländern in der Primarstufe installiert worden. Seit 2007 werden Vergleichsarbeiten in der Primarstufe einheitlich gegen Ende der dritten Jahrgangsstufe geschrieben, sodass die Schulen eine Rückmeldung darüber erhalten, wie weit ihre Schülerinnen und Schüler von den durch die KMK (KMK, 2004a-b) in den Bildungsstandards veröffentlichten Kompetenzerwartungen am Ende der Jahrgangsstufe vier entfernt sind. Neben dem Angebot des IQB bieten die Webseiten der Projektgruppe VERA an der Universität Koblenz-Landau, welche derzeit für acht Bundesländer ein gemeinsames Rückmeldesystem koordiniert, Beispielaufgaben und zahlreiche Materialien aus den Durchführungen 2004-2009 zum Download an (http://www.uni-landau.de/vera/).

2.4.1.1 Kompetenzskalen im Fach Deutsch bei VERA 3

Kompetenzbereiche im Fach Deutsch

„Sprache ist Träger von Sinn und Überlieferung, Schlüssel zum Welt- und Selbstverständnis und Mittel zwischenmenschlicher Verständigung. Sie hat grundlegende Bedeutung für die kognitive, emotionale und soziale Entwicklung der Kinder“, heißt es in den Bildungsstandards für die Primarstufe im Fach Deutsch (KMK, 2004a).

Für die Entwicklung von Kompetenzstufenmodellen für VERA wurden theoretische und empirische Erkenntnisse, beispielsweise aus den IGLU-Erhebungen (Internationale Grundschul-Lese-Untersuchung; Bos et al., 2003, 2007), LAU (Aspekte der Lernausgangslage und Lernentwicklung; Lehmann & Peek, 1997) und DESI (Deutsch Englisch Schülerleistungen International; DESI-Konsortium, 2008) aufgearbeitet. Die resultierenden Aufgabenmengen für das Fach Deutsch operationalisiert aus den Bildungsstandards die Kompetenzbereiche

- Leseverstehen,
- Hörverstehen,
- Sprache und Sprachgebrauch untersuchen,
- Schreiben und
- Orthografie.

Die Durchführung der VERA-Tests umfasst zwei Testtage, wobei der Kompetenzbereich Leseverstehen jeweils am ersten Testtag, einer der weiteren Kompetenzbereiche am zweiten Testtag erfasst wird.

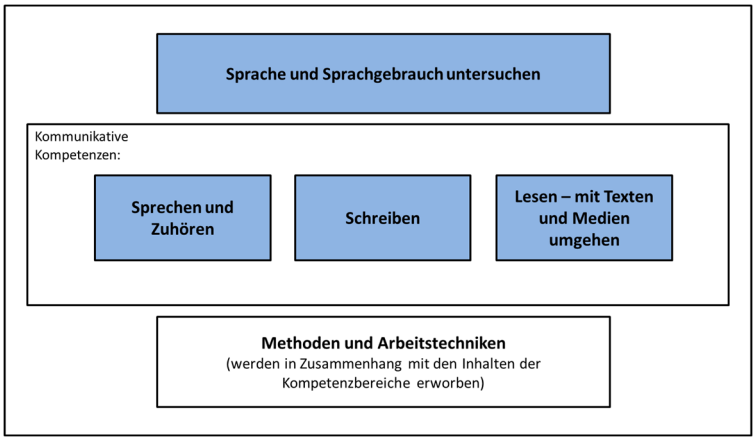


Abbildung 6: Kompetenzbereiche der Bildungsstandards für die Primarstufe im Fach Deutsch
Hinweis: Blau unterlegt jene Kompetenzbereiche, für die Kompetenzstufenmodelle (teilweise in Bezug auf Subkomponenten) entwickelt wurden.

Inhalte der Kompetenzbereiche und Aufgabenformate

Rechtschreibung spielt außerhalb der Kompetenzbereiche „Schreiben" und „Orthografie" eine untergeordnete Rolle. Entsprechend kommen in den anderen Kompetenzbereichen zumeist geschlossene Aufgabenformate wie Zuordnungs- oder Multiple-Choice-Aufgaben zum Einsatz, die geringe Textproduktion beinhalten. In dieser Entscheidung spiegelt sich das Bestreben wider, Teilleistungsbereiche der Fächer valide abzubilden. Die differenzierte und objektive Erfassung von Schreibleistungen im Rahmen von Vergleichsarbeiten stellt nach wie vor eine große Herausforderung dar. Die Schülerinnen und Schüler erhalten hier eine offene, aber klar umschriebene Aufgabenstellung zur Formulierung eines Textes. Die Auswertung findet anhand von Indikatoren statt, welche die Erfüllung von text- bzw. aufgabenspezifischen und allgemeinen Kompetenzmerkmalen (Anzahl der Wörter, Anzahl richtig geschriebener Wörter, Morphematik, Syntax und Textverknüpfungen) erfragen und somit einen indirekten Schluss auf Schreibkompetenzen der Schülerinnen und Schüler zulassen. Tabelle 3 zeigt ein Beispiel der VERA 3-Erhebung 2009 für einen solchen Indikator des Kompetenzbereichs „Schreiben" (Projekt VERA, 2009).

Kriterium	Punktevergabe	Beschreibung	Beispiel
nicht erfüllt	0	• Es gibt keine konsistente Perspektive. • Im Text werden mindestens 3 verschiedene Perspektiven eingenommen, z.B. Ich-Wir, Sie-Du, Ihr-Die Eltern. • Ob ein Wechsel zur neutralen Perspektive (man) einen Bruch darstellt, hängt von der Perspektive des restlichen Testes ab.	„Es wäre nett, dass **SIE** und die Kinder kommen. Natürlich könnt **IHR** auch Bekannte einladen […] Essen und Trinken sollte MAN mitbringen…"
teilweise erfüllt	1	• Die Adressaten werden nicht durchgängig angesprochen • Es werden zwei verschiedene Perspektiven eingenommen, gleichgültig wie oft zwischen ihnen gewechselt wird • Ob ein Wechsel zur neutralen Perspektive (man) einen Bruch darstellt, hängt von der Perspektive des restlichen Textes ab.	„Bei unserem Programm pusten WIR Luftballons auf, dann essen WIR, dann basteln **WIR** Piratensachen, dann hört **MAN** eine Geschichte, dann singen **WIR**."
voll erfüllt	2	• Die Adressaten werden durchgängig angesprochen. • Die Schreibperspektive (Schüler, Klasse, Lehrkraft) wird konsequent beibehalten. Dabei kann es sich auch um eine neutrale Perspektive handeln, in der niemand direkt angesprochen wird. Ob ein Wechsel zur neutralen Perspektive (man) einen Bruch darstellt, hängt von der Perspektive des restlichen Textes ab.	„**WIR** haben ein tolles Programm vorbereitet mit vielen Spielen. **MAN** kann dabei auch Preise gewinnen und Spaß haben."

Tabelle 3: Beispiel für einen Indikator des Kompetenzbereichs „Schreiben" im Fach Deutsch aus der VERA 3-Erhebung 2009 (Projekt VERA, 2009)

Im Folgenden wird beispielhaft der Kompetenzbereich „Sprache und Sprachgebrauch untersuchen“ beschrieben.

Beispiel:

Beispiel Kompetenzbereich „Sprache und Sprachgebrauch untersuchen“

„Sprache und Sprachgebrauch untersuchen“ findet sich als Kompetenzbereich sowohl in den Bildungsstandards der Primarstufe als auch in den Standards für den Hauptschul- und mittleren Bildungsabschluss der Sekundarstufe I. Er überspannt die kommunikativen Kompetenzen Lesen, Hören, Schreiben und Sprechen, für die grundlegende Kompetenzen im Bereich „Sprache und Sprachgebrauch untersuchen“ eine Voraussetzung darstellen. Der Inhalt der Kompetenzen in der Primarstufe bezieht sich auf den bewussten Umgang mit und die bewusste Anwendung von Sprache sowie auf ein grundlegendes Wissen über grammatische Strukturen und die Bedeutung von Wörtern, Sätzen und Texten. Aufgaben des Kompetenzbereichs sollen Schülerinnen und Schülern anregen, Sprache in Bezug auf ihre inhaltliche Dimensionen und Verwendungszusammenhänge zu untersuchen. Dabei kann die Grundschule an frühe Spracherfahrungen anknüpfen, denn schon vor der Einschulung haben die Kinder eigenaktiv Wissen über Sprache erworben, und auch außerhalb des Unterrichts wird dieses Wissen parallel zur Schule weiterentwickelt. Es ist dabei zielführend, zwischen deklarativem (Wissen über Sprache) und prozeduralem Wissen (Fertigkeiten im Umgang mit Sprache) zu unterscheiden. Beide Wissensbereiche müssen nicht übereinstimmen, sondern können bei den Kindern sehr unterschiedlich stark ausgeprägt sein. Tabelle 4 zeigt beispielhaft die Inhalte der Teilkompetenzen „An Wörtern, Sätzen, Texten arbeiten“ sowie „Grundlegende sprachliche Strukturen und Begriffe kennen und verwenden“, die bei VERA ins Zentrum des Kompetenzbereichs gerückt worden sind - nicht zuletzt, da sie sich derzeit am besten operationalisieren lassen.

Teilkompetenz		Inhalte
an Wörtern, Sätzen, Texten arbeiten, z.B.		- Wörter strukturieren und Möglichkeiten der Wortbildung kennen, - Wörter sammeln und ordnen, - sprachliche Operationen nutzen: umstellen, ersetzen, ergänzen, weglassen, - die Textproduktion und das Textverständnis durch die Anwendung von sprachlichen Operationen unterstützen, - mit Sprache experimentell und spielerisch umgehen.
grundlegende sprachliche Strukturen und Begriffe kennen und verwenden, z.B.	**Wort**	- Buchstabe, Laut, Selbstlaut, Mitlaut, Umlaut, Silbe, Alphabet - Wortfamilie, Wortstamm, Wortbaustein; Wortfeld; Wortart - Nomen: Einzahl, Mehrzahl, Fall, Geschlecht - Verb: Grundform, gebeugte Form - Zeitformen: Gegenwart, Vergangenheitsformen - Artikel: bestimmter Artikel, unbestimmter Artikel - Adjektiv: Grundform, Vergleichsstufen - Pronomen - andere Wörter (alle hier nicht kategorisierten Wörter gehören zu dieser Restkategorie)
	Satz	- Satzzeichen: Punkt, Komma, Fragezeichen, Ausrufezeichen, - Doppelpunkt, Redezeichen - Satzart: Aussage-, Frage-, Ausrufesatz - wörtliche Rede - Subjekt - Prädikat/Satzkern - Ergänzungen: Satzglied; einteilige, mehrteilige Ergänzung - Vergangenheit, Gegenwart, Zukunft (als Zeitstufen)

Tabelle 4: Inhalte der bei VERA berücksichtigten Teilkompetenzen „An Wörtern, Sätzen, Texten arbeiten“ und „Grundlegende sprachliche Strukturen und Begriffe kennen und verwenden“

Die Aufgaben zu diesem Kompetenzbereich beziehen sich auf Semantik und Stil, Morphologie, Syntax sowie Regelwissen und Rechtschreibung. Betont werden die Alltagsnähe der Aufgaben und die (implizite) Berücksichtigung grammatikalischer Begrifflichkeiten bei deren Konstruktion. Die Aufgabenformate sind dabei breit gefächert und beinhalteten geschlossene Formate wie Multiple-Choice, Ergänzungs-, Zuordnungs- und Unterstreichungsaufgaben genauso wie halboffene Formate (vgl. Projekt VERA, 2008; Isaac, Metzeld & Eichler, 2009).

2.4.1.2 Kompetenzskalen im Fach Mathematik bei VERA 3

Kompetenzen in der Primarstufen-mathematik

In der Grundschule soll der Mathematikunterricht entsprechend den Vorstellungen, wie sie in den Bildungsstandards dargelegt werden, mathematische Alltagserfahrungen der Schülerinnen und Schüler aufnehmen und weiterentwickeln (KMK, 2004b). Aus ihnen sollen sich frühe Kompetenzen entfalten, die eine Grundlage für die Auseinandersetzung mit mathematischen Anforderungen in späteren Lebenssituationen (inklusive der Schule) bilden. Darüber hinaus betonen die Bildungsstandards auch die Bedeutung einer grundlegenden positiven Einstellung gegenüber der Mathematik, die bereits in der Grundschule bestmöglich gefördert werden soll.

Die Bildungsstandards der Primarstufe für das Fach Mathematik beziehen sich nur indirekt auf Sachgebiete des Mathematikunterrichts der Grundschule (Arithmetik, Geometrie, Größen und Sachrechnen). Vielmehr werden allgemeine und inhaltsbezogene mathematische Kompetenzen abgegrenzt.

Allgemeine mathematische Kompetenzen

Allgemeine mathematische Kompetenzen beziehen sich auf die Art und Weise, wie Mathematik betrieben wird, und werden in der Auseinandersetzung mit Mathematik erworben. Die Ausprägung der jeweiligen Kompetenzen ist damit abhängig von der Quantität und Qualität der Unterrichtsgelegenheiten, in denen diese erprobt werden können. Die Bildungsstandards unterscheiden sechs Kompetenzbereiche (vgl. auch die Umschreibung der mathematischen Kompetenzen bei VERA 8 unten):

1. Technische Grundfertigkeiten
2. Problemlösen
3. Kommunizieren
4. Argumentieren
5. Modellieren
6. Darstellen

Inhaltsbezogene mathematische Kompetenzen

Inhaltsbezogene mathematische Kompetenzen beziehen sich dagegen über die Grenzen der (curricularen) Sachgebiete des Mathematikunterrichts hinweg auf grundlegende mathematische Konzepte, die in fünf mathematische Leitideen gegliedert sind:

1. Zahlen und Operationen
2. Raum und Form
3. Muster und Strukturen
4. Größen und Messen
5. Daten, Häufigkeit und Wahrscheinlichkeit

Bereits in der Primarstufen-Mathematik greifen die Bildungsstandards damit wesentliche Ideen von Freudenthal (1977; 1983) auf, wonach der verständnisvolle Umgang mit Mathematik und ihre flexible Anwendung in verschiedenartigen Kontexten im Zentrum mathematischen Denkens stehen. In der Primarstufe kann diese Entwicklung in Gang gesetzt werden, indem Alltagserfahrungen der Schülerinnen und Schüler aufgegriffen werden.

Das vom IQB entwickelte Kompetenzstufenmodell beinhaltet alle in den Bildungsstandards ausgewiesenen mathematischen Leitideen und ist in fünf Kompetenzstufen aufgebaut, die von der Beschreibung basaler Kompetenzen bis hin zum elaborierten Umgangs mit mathematischen Inhalten in der Primarstufe reichen (vgl. IQB, 2010; vgl. auch Reiss & Winkelmann 2009). Zentrale Kompetenzanforderungen dieser Stufen sind in Tabelle 5 dargestellt (IQB, 2010). Für die Vergleichsarbeiten in der Grundschule wurden in den letzten Jahren jeweils zwei oder drei mathematische Leitideen ausgewählt, für die eine kompetenzbezogene Rückmeldung erfolgte.

Anforderungsbereiche

In den Anforderungsbereichen „Reproduzieren“, „Zusammenhänge erkennen“ und „Verallgemeinern und Reflektieren“ wird die (kognitive) Komplexität und Qualität der Kompetenzausprägung in den Bildungsstandards grob differenziert. Eine Anbindung an diese Anforderungsbereiche ist bei VERA allerdings bisher nicht vorgenommen worden.

Aufgabenformate

Zumeist wird auf geschlossene Aufgabenformate zurückgegriffen, insbesondere auf Ergänzungs- und Multiple-Choice-Aufgaben, die adressatengerecht häufig durch mathematische Abbildungen als Stimulus eingeleitet werden.

2.4.2 Vergleichsarbeiten in der achten Jahrgangsstufe (VERA 8)

Aufgabenentwicklung bei VERA 8

Das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) hatte im Zuge der Normierung der Bildungsstandards eine umfangreiche Aufgabenentwicklung vorangetrieben, aus der seit 2009 auch die Aufgaben für die achte Jahrgangsstufe hervorgehen. Die Kompetenzen werden auf fünf gleich großen Stufen beschrieben, wobei die Zuordnung zu Stufe 2 im Sinne des Erreichens der Mindeststandards (vgl. Kapitel 3.2 im Studienbrief Bildungsmonitoring) diskutiert wird.

Aufgabenzusammenstellungen in drei Schwierigkeitsstufen

Um der relativ großen Leistungsstreuung in Deutschland gerecht zu werden, werden Aufgabensätze in drei Schwierigkeitsstufen (Testheftversionen) entwickelt (vgl. http://www.iqb.hu-berlin.de/vera/faq). Welche dieser Schwierigkeitsstufen in welcher Schulart eingesetzt wird, bestimmen die Länder. Zumeist erhalten aber alle Schülerinnen und Schüler einer Klasse dieselbe Testheftversion. Wie in Kapitel 2.3 erläutert, ist es möglich, Testergebnisse dieser unterschiedlich schwierigen Testhefte zu vergleichen und damit auch die erworbenen Kompetenzen in Relation zu setzen.

Stufe	zentrale Anforderungsmerkmale
I a: Technische Grundlagen 1	Die Grundaufgaben des kleinen Einspluseins und Einmaleins werden beherrscht und genutzt, wenn die Aufgabenstellungen keine besonderen Schwierigkeiten aufweisen. Klar strukturierten Diagrammen, Schaubildern und Tabellen mit Bezug zur Lebenswirklichkeit können relevante Daten entnommen werden.
I b: Technische Grundlagen 2	Einfache mathematische Begriffe und Prozeduren sind bekannt und können in einem innermathematischen Kontext bzw. in einem aus dem Alltag vertrauten oder gut geübten Kontext korrekt reproduziert werden. Kleine Zahlen können in Bezug auf ihre Größe verglichen werden, Zahldarstellungen in Stellentafeln werden sicher gelesen. Auch die schwierigeren Einmaleinsaufgaben werden gelöst. Numerisches Wissen wird in einfachen Alltagssituationen angewendet. Einfache Reihungen werden erkannt und fortgesetzt. Insbesondere werden grundlegende Begriffe der ebenen Geometrie und gängige Repräsentanten standardisierter Einheiten richtig verwendet.
II: Einfache Anwendungen von Grundlagenwissen	Hier geht es um Routineprozeduren in einem klar strukturierten Kontext: Die Struktur des Dezimalsystems wird genutzt, Gesetzmäßigkeiten werden erkannt und bei der Fortsetzung einfacher Zahlenfolgen, beim strukturierten Zählen und systematischen Probieren berücksichtigt. Aufgaben zur Addition, Subtraktion und Multiplikation werden halbschriftlich und schriftlich durchgeführt, Überschlagsrechnungen werden durchgeführt. Insbesondere können in diesem Zusammenhang einfache Sachaufgaben gelöst werden. Aus dem Alltag vertraute proportionale Zuordnungen werden erkannt und angewendet. Bei einfachem Zahlenmaterial wird das Umwandeln von Größen in gegebene Einheiten auch bei gemischten Größenangaben durchgeführt. Grundbegriffe der räumlichen Geometrie werden korrekt verwendet, wenn diese einen Bezug zum Alltag haben. Räumliche Beziehungen werden zur Lösung einfacher Probleme genutzt. Wesentliche Grundbegriffe aus dem Umfeld von Zufall und Wahrscheinlichkeit werden korrekt verwendet („sicher“, „unmöglich“, „wahrscheinlich“).
III: Erkennen und Nutzen von Zusammenhängen	Potenzielle Zusammenhänge werden erkannt und in einem vertrauten (mathematischen und sachbezogenen) Kontext genutzt. Das erlernte Wissen kann auf dieser Stufe flexibel in unterschiedlichen Problemstellungen genutzt werden, die einem vertrauten Kontext zuzuordnen sind. Insbesondere wird mit Zahlen und Operationen im curricularen Umfang sicher umgegangen. Überschlagsrechnungen werden auch bei großen Zahlen sicher durchgeführt. Strukturelle Aspekte werden bei gut geübten Inhalten gesehen und können kommuniziert werden. Das betrifft auch Inhalte der Geometrie, wobei etwa zwischen verschiedenen Darstellungsformen einer Figur vermittelt werden kann. Einfache Sachsituationen werden modelliert und die damit verbundenen Problemstellungen gelöst. Daten und Informationen können in bekanntem Kontext flexibel dargestellt werden. Bei nicht allzu komplexen Zufallsexperimenten werden Gewinnchancen korrekt eingeschätzt und begründet.
IV: Sicheres und flexibles Anwenden von begrifflichem Wissen und Prozeduren in curricularem Umfang	Auch in einem wenig vertrauten Kontext wird mathematisches Wissen sicher angewendet. Eigene Vorgehensweisen werden korrekt beschrieben, die Lösungswege anderer Kinder werden verstanden und reflektiert. Das Rechnen wird im curricularen Umfang in allen Varianten sicher beherrscht. Begriffe der ebenen und räumlichen Geometrie werden flexibel verwendet. Zahldarstellungen in Stellenwerttafeln können auch bei sehr großen Zahlen nach Vorschrift selbstständig manipuliert und systematisch verändert werden. Das Rechnen mit Größen ist sicher und flexibel und umfasst insbesondere Nähe-

	rungsrechnungen und Überschlagsrechnungen. Informationen aus unterschiedlichen Quellen können in einen Zusammenhang gestellt und in Modellierungsaufgaben selbstständig verwendet und manipuliert werden.
V: Modellierung komplexer Probleme unter selbstständiger Entwicklung geeigneter Strategien	Mathematische Problemstellungen werden auch in einem unbekannten Kontext angemessen, sicher und flexibel bearbeitet. Dabei werden geeignete Strategien, sinnvolle Bewertungen oder Verallgemeinerungen auf hohem Niveau geleistet. Umfangreiches curricular verankertes Wissen wird in ungewohnten Situationen flexibel genutzt. Das Vorgehen kann sicher und nachvollziehbar kommuniziert und begründet werden. Komplexe Sachsituationen werden modelliert und bearbeitet, wobei besondere Schwierigkeiten wie die Verwendung von Tabellen, der Umgang mit zusammengesetzten Größen oder das Rechnen mit Zahlen in Kommaschreibweise auftreten können. Es können auch ungewohnte funktionale Zusammenhänge analysiert und genutzt werden. Die Lösung von Aufgaben kann ein hohes Maß an räumlichem Denken oder entsprechende analytische Fähigkeiten voraussetzen.

Tabelle 5: Zentrale Anforderungsmerkmale der Kompetenzstufen im Fach Mathematik aus der VERA 3-Erhebung 2010 (IQB, 2010)

Die folgende Darstellung der Kompetenzstufenmodelle für Deutsch, Englisch und Mathematik kann durch Aufgabenbeispiele illustriert werden (http://www.iqb.hu-berlin.de/vera/aufgaben).

2.4.2.1 Kompetenzskalen im Fach Deutsch bei VERA 8

Kompetenzskalen im Fach Deutsch bei VERA 8

Kompetenz in der deutschen Sprache geht nach den Vorstellungen der KMK weit über eine reine Verstehens- und Verständigungskompetenz hinaus (KMK, 2003a; 2004c). Die in den Bildungsstandards beschriebenen Erwartungen sollen vielmehr als Voraussetzungen dafür dienen, dass Schülerinnen und Schüler Orientierungs- und Handlungswissen gewinnen, etwa wenn es darum geht, „kritische Distanz zwischen Lebenswirklichkeit und den in Literatur und Medien dargestellten virtuellen Welten“ zu entwickeln, sich die „Bedeutung des Reichtums kultureller, sprachlicher, literarischer und medialer Vielfalt“ zu vergegenwärtigen oder im direkten Austausch und in der Auseinandersetzung mit kulturellen Traditionen Fremdverstehen und Toleranz zu fördern. Die Vorlage der KMK für die Entwicklung von Kompetenzstufenmodellen zur Untersuchung des Lernstands im Fach Deutsch war also umfangreich.

Im Fach Deutsch gingen weitreichende Überlegungen zur Aufgabenentwicklung und zu den Kompetenzstufenmodellen aus dem Projekt „Evaluation der Standards Deutsch für die Sekundarstufe I“ (ESDeS I; Federführung: Prof. Dr. Bremerich-Vos, Universität Duisburg-Essen) hervor. Darüber hinaus lagen Erfahrungen aus den Large-Scale-Assessments des Bildungsmonitorings, insbesondere aus PISA (Kompetenzbereich Lesen, vor allem PISA 2000; Artelt, Stanat, Schneider & Schiefele, 2001) und DESI (DESI-Konsortium, 2008) vor. Aktuell liegen für den mittleren Schulabschluss Kompetenzstufenmodelle für folgende Kompetenzbereiche vor (vgl. die Webseite des IQB: http://www.iqb.hu-berlin.de/bista?reg=r_4; 22.06.2011):

- Lesen
- Orthografie
- Sprechen und Zuhören (hier zunächst allerdings nur die Subkomponente „verstehend zuhören“).

Die konstruierten Aufgaben „tragen den aktuellen Erkenntnissen der Fachdidaktik und der Psychometrie Rechnung, um einerseits der Bedeutung des Deutschunterrichts hinsichtlich seines Beitrags zur sprachlichen, literarischen und medialen Bildung der Schülerinnen und Schüler gerecht zu werden und andererseits empirisch verlässliche Messinstrumente zu erhalten“ (IQB, ohne Jahr). Dies wird hier beispielhaft für den Kompetenzbereich Lesen beschrieben (vgl. IQB, 2009).

Beispiel: Kompetenzbeschreibungen im Lesen – Komponente „Lesen – mit Texten und Medien umgehen“

Beispiel Kompetenzbereich „Lesen“

Leseverstehen beschreibt die Fähigkeit, als Schrift- oder Bildzeichen kodierte Informationen in schriftlichen Dokumenten zu verstehen (Schnotz & Dutke, 2004). Diese Informationen können als Texte, aber auch in Form von Bildern, Diagrammen oder Tabellen dargestellt sein. Leseverstehen beinhaltet kognitive Prozesse, bei der die Lesenden aktiv eine mentale Repräsentation der Textinhalte vornehmen und ihnen Bedeutung verleihen. Zudem ist für die Ausprägung von Lesekompetenz die Motivation zum Lesen

relevant, von der angenommen wird, dass sie über die Lesemenge und den Einsatz von Lesestrategien die Kompetenzausprägung beeinflusst (Möller & Schiefele, 2004).

In den Bildungsstandards für das Fach Deutsch besitzt die Lesekompetenz – zweifellos auch unter dem Eindruck von PISA – eine hohe Bedeutung: Lesekompetenz wird darin als Schlüsselkompetenz zur Teilhabe am gesellschaftlichen Leben angesehen. Einige der dort beschriebenen Komponenten (z.B. Anschlusskommunikation) sind diagnostisch nur schwer zu erfassen, insbesondere im Rahmen von groß angelegten Tests im Klassenverband.

Die bisher entstandenen Kompetenzstufenmodelle berücksichtigen vornehmlich das Leseverstehen bei literarischen Texten sowie Sach- und Gebrauchstexten (vgl. die Kompetenzanforderungen in Tabelle 6). Es wurde hier also weitestgehend den konzeptionellen Vorarbeiten aus Studien des Bildungsmonitoring wie PISA oder IGLU (PIRLS) gefolgt, bei denen ebenfalls nach dem Textgenre differenziert wird. Daneben sind Leseprozesse (Informationen ermitteln vs. textbezogenes Interpretieren vs. Reflektieren und Bewerten) und das Textformat (kontinuierliche vs. diskontinuierliche Texte) als relevante differenzierende Merkmale der Lesekompetenz erfasst worden. Alle konstruierten Aufgaben lassen sich dennoch auf einer gemeinsamen Kompetenzskala abbilden, da Schülerinnen und Schülern in diesen verschiedenen Merkmalsbereichen sehr ähnliche Ergebnisse erbringen.

Bei der Erfassung von Lesekompetenz in den Vergleichsarbeiten stehen die folgenden Aspekte im Vordergrund:

- wesentliche Elemente eines Textes erfassen
- wesentliche Fachbegriffe zur Erschließung von Literatur kennen und anwenden
- eigene Deutungen des Textes entwickeln
- Informationen zielgerichtet entnehmen, ordnen (...)
- Medien verstehen und nutzen
- zwischen eigener Wirklichkeit und virtuellen Welten in Medien unterscheiden.

Grundlegende Verfahren

- Lesetechniken und Strategien zum Leseverstehen kennen und anwenden
- über grundlegende Lesefertigkeiten verfügen: flüssig, sinnbezogen, überfliegend, selektiv
- die eigenen Leseziele kennen
- Vorwissen und neue Informationen unterscheiden
- Wortbedeutungen klären
- Lesehilfen nutzen: z.B. Textsorte, Aufbau, Überschrift, Illustration, Layout
- Verfahren zur Textstrukturierung kennen und nutzen: Inhalte zusammenfassen, Zwischenüberschriften formulieren, wesentliche Textstellen kennzeichnen, Bezüge zwischen Textstellen herstellen, Fragen aus dem Text ableiten und beantworten
- Verfahren zur Textaufnahme kennen und nutzen: Aussagen erklären, Stichwörter formulieren, Texte und Textabschnitte zusammenfassen

Literarische Texte verstehen und nutzen

- aktuelle und klassische Werke der Jugendliteratur und altersangemessene Texte bedeutender Autorinnen und Autoren kennen
- epische, lyrische, dramatische Texte unterscheiden und wesentliche Merkmale kennen, insbesondere epische Kleinformen, Erzählung, Kurzgeschichte, Gedichte
- an einem repräsentativen Beispiel Zusammenhänge zwischen Text, Entstehungszeit und Leben des Autors/der Autorin herstellen
- zentrale Aussagen erschließen
- wesentliche Elemente eines Textes erfassen: Figuren, Raum- und Zeitdarstellung, Konfliktverlauf
- Handlung und Verhaltensweisen beschreiben und werten
- wesentliche Fachbegriffe zur Erschließung von Literatur kennen und anwenden: Autor, Erzähler, Monolog, Dialog, Reim
- grundlegende Gestaltungsmittel erkennen und ihre Wirkungen einschätzen: z.B. Wortwahl, Wiederholung, sprachliche Bilder
- untersuchende und produktive Methoden kennen und anwenden: z.B. Texte vergleichen, weiterschreiben, Paralleltext verfassen, szenische Umsetzung
- eigene Deutungen des Textes entwickeln, mit anderen darüber sprechen und am Text belegen

Sach- und Gebrauchstexte verstehen und nutzen
- verschiedene Textfunktionen und Textsorten unterscheiden: informieren (z.B. Lexikontext), appellieren (z.B. Werbetext), regulieren (z.B. Jugendschutzgesetz, Arbeitsvertrag), instruieren (z.B. Bedienungsanleitung) - Informationen zielgerichtet entnehmen, ordnen, prüfen und ergänzen - nichtlineare Texte (auch im Zusammenhang mit linearen Texten) auswerten: z.B. Schaubilder, - Intention(en) eines Textes erkennen - aus Sach- und Gebrauchstexten begründete Schlussfolgerungen ziehen - Information und Wertung in Texten unterscheiden: z.B. in Zeitungen
Medien verstehen und nutzen
- Informations- und Unterhaltungsfunktion unterscheiden: z.B. im Internet - wesentliche Darstellungsmittel eines Mediums und deren Wirkungen kennen und einschätzen - Intentionen und Wirkungen ausgewählter Medieninhalte erkennen und bewerten: z. B. Fernsehserie - Lebenswirklichkeit von Realitätsdarstellungen und der Darstellung fiktionaler Welten in Medien unterscheiden - Informationen zu einem Thema/Problem in unterschiedlichen Medien suchen, vergleichen, auswählen und bewerten - Medien für die eigene Produktion kreativ nutzen
Methoden und Arbeitstechniken
- mit Nachschlagewerken umgehen können - recherchieren - zitieren, Quellen angeben - Wesentliches markieren - Stichwörter formulieren - Texte gliedern und Teilüberschriften finden - Inhalte mit eigenen Worten zusammenfassend wiedergeben - Arbeitsergebnisse zielgerichtet und sachbezogen präsentieren z.B. mit Folie, Plakat, PC

Tabelle 6: Teilkompetenzen des Kompetenzbereichs „Lesen – mit Texten und Medien umgehen“ im Fach Deutsch in den Bildungsstandards für den Hauptschulabschluss

Aufgabenformate Leseverstehen

Überwiegend wird dabei auf authentische Texte zurückgegriffen, die sich in Länge und Komplexität deutlich unterscheiden können. Als Aufgabenformate kommen zumeist Multiple-Choice-, Wahr/Falsch- und Zuordnungsaufgaben zum Einsatz, teilweise sind Kurzantworten zu geben. Alle Aufgaben werden dichotom ausgewertet, es gibt also keine „Zwischen-“ oder „Teillösungen“, und auch nicht bearbeitete Items werden als „falsch“ gewertet. Diese Art der Auswertung unterscheidet sich damit beispielsweise von differenzierteren Bewertungsschemata in Klassenarbeiten, bei deren Korrektur das Prinzip gilt, jede erbrachte Leistung durch Teilpunkte zu würdigen. Vergleichsarbeiten sind dagegen so konzipiert, dass mit dem Ergebnis ausgelotet werden kann, auf welcher Kompetenzstufe sich die Schülerinnen und Schüler mit einer gewissen Wahrscheinlichkeit befinden. Damit Schülern z. B. eine hohe Kompetenzstufe zugewiesen werden kann, müssen sie in der Lage sein, mit einer hinreichenden Sicherheit auch schwierige, komplexe Aufgaben vollständig zu lösen. Um bei diesen Aufgaben zu Teillösungen zu gelangen, würde eine geringere Kompetenz genügen, welche sich aber bereits durch die Bearbeitung von leichten Aufgaben nachweisen lässt.

2.4.2.2 Kompetenzskalen im Fach Englisch bei VERA 8

Kompetenzskalen im Fach Englisch bei VERA 8

Tenorth (2001) stellte fest, dass der Beherrschung der englischen Sprache als lingua franca eine besondere Bedeutung unter den Neu- und Altsprachen zukommt. Sprachliche Kompetenz im Englischen ist längst eine „notwendige Voraussetzung für die Teilnahme an der Leistungsgesellschaft“. So verwundert es nicht, dass sich die KMK mit dem Auftrag für die Schulleistungsstudie DESI (Deutsch-Englisch-Schülerleistungen-International; DESI-Konsortium, 2008; vgl. Studienbrief 3) bemühte, eine Lücke in der Gesamtstrategie zum Bildungsmonitoring zu schließen, und Vergleichsarbeiten in Englisch als erster Fremdsprache der meisten Schülerinnen und Schüler – neben Deutsch und Mathematik – in nahezu allen Bundesländern implementiert wurden.

Bildungsstandards in der ersten Fremdsprache

Die Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) greifen auf den Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER; Europarat, 2001; vgl. Tabelle 7) zurück, in dem produktive und rezeptive Kompetenzen der Fremdsprachenverwendung auf verschiedenen Dimensionen definiert sind. Diese Kompetenzen sollen die Handlungsfähigkeit von Personen im öffentlichen, privaten und beruflichen Leben gewährleisten.

Leistungsniveaus des GER

Für das Fach Englisch definieren die nationalen Bildungsstandards Anforderungen in Bezug auf funktionale kommunikative, interkulturelle und methodische Kompetenzen (Abbildung 7; vgl. KMK, 2003b; 2004d). Allerdings lassen sich vor allem kommunikative Kompetenzen in der englischen Sprache im Rahmen von Vergleichsarbeiten sinnvoll erfassen, sodass sich die bisher entwickelten Kompetenzstufenmodelle auf die Kompetenzbereiche Hörverstehen und Leseverstehen beschränken (http://www.iqb.hu-berlin.de/bista?reg=r_4; 22.06.2011).

Im GER sind in Form von Deskriptoren erwartbare Leistungen definiert, die insgesamt sechs Leistungsniveaus (A1, A2, B1, B2, C1, C2) unterscheiden. Durch das IQB wurden Testaufgaben entwickelt, welche die Niveaus A1 bis C1 abdecken, sodass sich in den Modellen also insgesamt fünf Kompetenzstufen abgrenzen lassen. Das Niveau C2 wurde hingegen ausgelassen, da es sich hierbei um Anforderungen handelt, die in der Regel erst in der Sekundarstufe II Berücksichtigung finden. Die Bildungsstandards für den Hauptschulabschluss orientieren sich im Kern an den Niveaus A2 und B1, die Bildungsstandards für den Mittleren Schulabschluss an den Niveaus B1 und B2. Tabelle 8 zeigt beispielhaft die Zuordnung von Kompetenzerwartungen zu den GER-Niveaus anhand des Kompetenzbereichs Hörverstehen. Die Aufgaben der Vergleichsarbeiten beinhalten ebenfalls eine Zuschreibung zu den Niveaustufen des GER.

Inhalte, Schwierigkeitsmerkmale und Operationalisierung der Standards für das Fach Englisch sowie Implikationen für Vergleichsarbeiten werden im Folgenden am Beispiel des Kompetenzbereiches „Hörverstehen“ dargestellt.

deutscher Schulabschluss	**GER-Niveau**		**Spezifikation des Niveaus**
Hauptschulabschluss	A:	Elementare Sprachverwendung	A1: Breakthrough A2: Waystage
mittlerer Bildungsabschluss	B:	Selbstständige Sprachverwendung	B1: Threshold B2: Vantage
Abitur	C:	Kompetente Sprachverwendung	C1: Effective Operational Proficiency C2: Mastery

Tabelle 7: Niveaus des gemeinsamen europäischen Referenzrahmens (GER)

Beispiel: Kompetenzbeschreibungen Englisch, Kompetenz „Hörverstehen“, mittlerer Bildungsabschluss

Beispiel Kompetenzbereich „Hörverstehen“

Bei der Erfassung von Hörverstehen wird der Stimulus „in Echtzeit“ akustisch dargeboten. Vor allem Besonderheiten in der sprachlichen Darbietung wie Dialekte und die Flüchtigkeit der Stimulusdarbietung machen dabei die Schwierigkeit des Verstehensprozesses aus (siehe unten). Die kognitiven Prozesse beim Hörverstehen reichen von der akustischen Wahrnehmung kurzzeitlich dargebotener Signale bis zur Sinngebung der akustischen Informationen durch die Integration von sprachlichem Wissen sowie Sach- und Weltwissen.

Hörverstehen ist bereits in der Erstsprache ein komplexer Vorgang; beim Erlernen einer Fremdsprache werden jedoch besonders hohe Anforderungen an das Hörverstehen gestellt. Dies liegt zum einen daran, dass Verstehenslücken beim Hörverstehen aufgrund der kurzzeitigen, flüchtigen Darbietung des akustischen Reizes nur noch sehr schwer gefüllt werden können. Zum anderen setzt Hörverstehen auch Vorwissen voraus. Dies bezieht sich nicht nur auf Vokabeln, sondern insbesondere auch auf Sinnzusammenhänge, die nur durch Wissen über Inhalte und kulturelle Besonderheiten hergestellt werden können. Fehlt dieses Wissen, so können Missverständnisse entstehen oder Texte nicht verstanden werden.

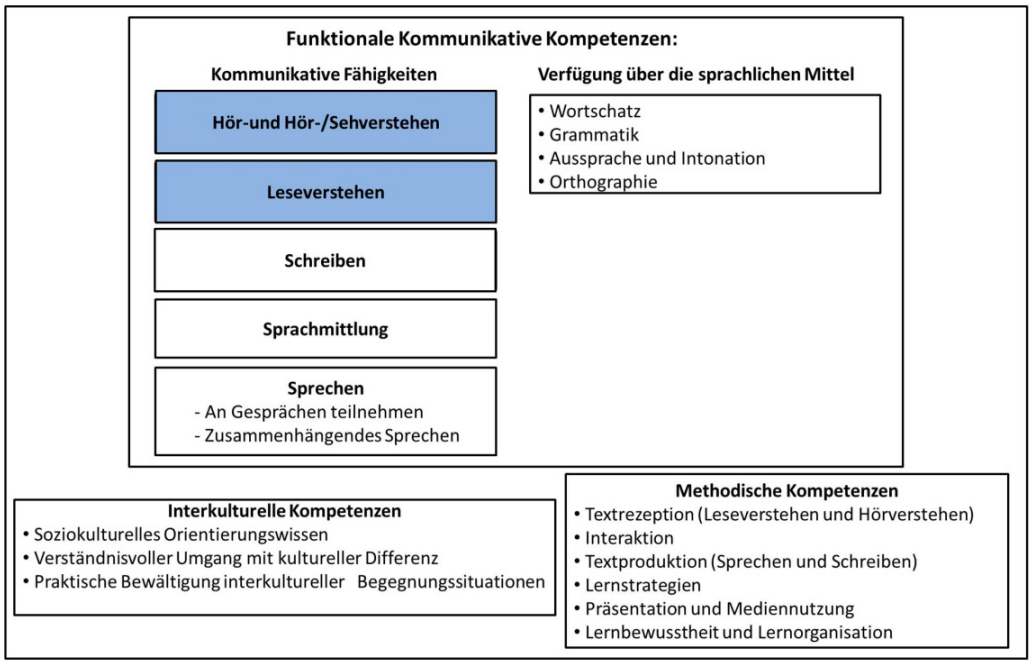


Abbildung 7: Funktional kommunikative, interkulturelle und methodische Kompetenzen der Bildungsstandards für den Hauptschul- und mittleren Bildungsabschluss in der ersten Fremdsprache
Hinweis: Blau unterlegt sind jene Kompetenzbereiche, für die durch das IQB Kompetenzstufenmodelle entwickelt wurden (http://www.iqb.hu-berlin.de/bista?reg=r_4; 22.06.2011).

Aber auch die Eigenschaften der Hörtexte und die Formulierung der Höraufgabe beeinflussen den Schwierigkeitsgrad des Hörverstehens. So haben sich als schwierigkeitsbestimmende Merkmale des Textes u. A. herausgestellt: die Länge des Textes, die Sprechgeschwindigkeit, die Anzahl und Aussprache der Sprecherinnen und Sprecher, Eigenheiten der verwendeten Sprache (z.B. Dialekte), die Textstruktur (Gliederung des Textes) und die lexikalische und grammatische Komplexität (Häufigkeit oder Abstraktionsgrad der Begriffe, Komplexität der Satzstrukturen).

Die Bildungsstandards betonen die Bedeutung von Hörverstehen als Teil kommunikativer Kompetenzen, die es ermöglichen, sich in der Fremdsprache „in Alltagssituationen und über lebenspraktische Angelegenheiten“ zu verständigen (KMK, 2003b). Hörverstehen besitzt aber auch Relevanz für die spätere berufliche Tätigkeit, etwa bei Telefonaten oder Bewerbungen im Ausland. Die Kompetenzbereiche des Hörverstehens beinhalten in den Bildungsstandards die Rezeption direkter Kommunikation zwischen Sprechern als auch mediale Kommunikation.

In der Formulierung der Bildungsstandards heißt es (KMK, 2003b; jeweiliges GER-Niveau in Klammern): „Die Schülerinnen und Schüler können unkomplizierte Sachinformationen über gewöhnliche alltags- oder berufsbezogene Themen verstehen und dabei die Hauptaussagen und Einzelinformationen erkennen, wenn in deutlich artikulierter Standardsprache gesprochen wird (B1+).

Die Schülerinnen und Schüler können

- im Allgemeinen den Hauptpunkten von längeren Gesprächen folgen, die in ihrer Gegenwart geführt werden (B1),
- Vorträge verstehen, wenn die Thematik vertraut und die Darstellung unkompliziert und klar strukturiert ist (B1+),
- Ankündigungen und Mitteilungen zu konkreten Themen verstehen, die in normaler Geschwindigkeit in Standardsprache gesprochen werden (B2),
- vielen Filmen folgen, deren Handlung im Wesentlichen durch Bild und Aktion getragen wird (B1),
- den Informationsgehalt der meisten Rundfunksendungen und Tonaufnahmen über Themen von persönlichem Interesse verstehen (B1+),
- das Wesentliche in vielen Fernsehsendungen zu Themen von persönlichem Interesse, z. B. Interviews, kurze Vorträge oder Nachrichtensendungen verstehen (B1+).“

Die Aufgaben zur Erfassung von Hörverstehen werden dahingehend konstruiert, dass sie möglichst authentische Texte aus englischsprachigen Ländern und unterschiedliche Textarten aufgreifen sowie eine Progression in den Leistungsanforderungen beinhalten. Anwendung finden vor allem Aufgabenformate wie Multiple Choice oder Multiple Matching, die eigene Textproduktion auf ein Mindestmaß reduzieren.

Bei der Auswertung spielen dementsprechend orthographische und grammatikalische Fehler nur eine untergeordnete Rolle.

A Elementare Sprachverwendung	
A1	Kann verstehen, wenn sehr langsam und sorgfältig gesprochen wird und wenn lange Pausen Zeit lassen, den Sinn zu erfassen.
A2	Versteht genug, um Bedürfnisse konkreter Art befriedigen zu können, sofern deutlich und langsam gesprochen wird. Kann Wendungen und Wörter verstehen, wenn es um Dinge von ganz unmittelbarer Bedeutung geht (z. B. ganz grundlegende Informationen zu Person, Familie, Einkaufen, Arbeit, nähere Umgebung) sofern deutlich und langsam gesprochen wird.
B Selbständige Sprachverwendung	
B1	Kann unkomplizierte Sachinformationen über gewöhnliche alltags- oder berufsbezogene Themen verstehen und dabei die Hauptaussagen und Einzelinformationen erkennen, sofern klar artikuliert und mit vertrautem Akzent gesprochen wird. Kann die Hauptpunkte verstehen, wenn in deutlich artikulierter Standardsprache über vertraute Dinge gesprochen wird, denen man normalerweise bei der Arbeit, in der Ausbildung oder der Freizeit begegnet; kann auch kurze Erzählungen verstehen.
B2	Kann im direkten Kontakt und in den Medien gesprochene Standardsprache verstehen, wenn es um vertraute oder auch um weniger vertraute Themen geht, wie man ihnen normalerweise im privaten, gesellschaftlichen, beruflichen Leben oder in der Ausbildung begegnet. Nur extreme Hintergrundgeräusche, unangemessene Diskursstrukturen oder starke Idiomatik beeinträchtigen das Verständnis. Kann die Hauptaussagen von inhaltlich und sprachlich komplexen Redebeiträgen zu konkreten und abstrakten Themen verstehen, wenn Standardsprache gesprochen wird; versteht auch Fachdiskussionen im eigenen Spezialgebiet. Kann längeren Redebeiträgen und komplexer Argumentation folgen, sofern die Thematik einigermaßen vertraut ist und der Rede- oder Gesprächsverlauf durch explizite Signale gekennzeichnet ist.
C Kompetente Sprachverwendung	
C1	Kann genug verstehen, um längeren Redebeiträgen über nicht vertraute abstrakte und komplexe Themen zu folgen, wenn auch gelegentlich Details bestätigt werden müssen, insbesondere bei fremdem Akzent. Kann ein breites Spektrum idiomatischer Wendungen und umgangssprachlicher Ausdrucksformen verstehen und Registerwechsel richtig beurteilen. Kann längeren Reden und Gesprächen folgen, auch wenn diese nicht klar strukturiert sind und wenn Zusammenhänge nicht explizit ausgedrückt sind.
C2	Hat keinerlei Schwierigkeiten, alle Arten gesprochener Sprache zu verstehen, sei dies live oder in den Medien, und zwar auch wenn schnell gesprochen wird, wie Muttersprachler dies tun.

Tabelle 8: Deskriptoren des GER für die Niveaus A1 bis C2 im Kompetenzbereich „Hörverstehen“ in der ersten Fremdsprache

2.4.2.3 Kompetenzskalen im Fach Mathematik bei VERA 8

Kompetenzskalen im Fach Mathematik bei VERA 8

Bei der Entwicklung von Aufgaben für das Fach Mathematik konnten die Beteiligten auf eine weitreichende Erfahrung aus den PISA-Studien (insbesondere PISA 2003; vgl. Studienbrief 3) zurückgreifen. Im Kontext der nationalen Ergänzungen der PISA-Studien in Deutschland waren eigene Aufgaben entwickelt worden, die der Aufgabenentwicklung für die nationalen Bildungsstandards im Fach Mathematik prägend vorausgingen. Charakteristisch für die Erfassung von mathematical literacy bei PISA ist die inhaltliche Differenzierung der Aufgaben nach allgemeinen mathematischen Kompetenzen und Leitideen (vgl. auch das Kompetenzstufenmodell für die Primarstufe oben). Für die Entwicklung der Aufgaben zur Überprüfung und Normierung der Bildungsstandards (KMK, 2003c; 2004e) wurde als zusätzliche Dimension der mathematische Anforderungsbereich spezifiziert. Diese Dimension war mehr oder weniger explizit ebenfalls bei PISA bereits herangezogen worden.

Die Aufgaben, welche die Bildungsstandards operationalisieren, füllen somit einen dreidimensionalen Raum, der fünf Leitideen, sechs Kompetenzen und drei Anforderungsbereiche abbildet (Abbildung 8, S. 26; vgl. Blum, Drüke-Noe, Hartung & Köller, 2006).

Leitideen

Das Konzept der Leitideen nimmt Überlegungen der Freudenthal-Didaktik (Freudenthal, 1977/1983) auf. Darin werden Phänomene einer mathematischen Betrachtungsweise beschrieben, welche den Stoffgebieten der Schulmathematik (Algebra, Arithmetik, Geometrie und Stochastik) vorausgehen und konzeptionell von diesen abzugrenzen sind. Die fünf *Leitideen* umfassen:

- Zahl,
- Messen,

- Raum und Form,
- Funktionaler Zusammenhang,
- Daten und Zufall.

Anforderungsbereiche

In den *Anforderungsbereichen* wird die Komplexität einer Aufgabe berücksichtigt. Diese ist nicht direkt auf die empirische Schwierigkeit übertragbar, spiegelt sich in der Regel darin aber wider. In den Bildungsstandards werden drei nach zunehmender Komplexität geordnete Anforderungsbereiche (AB) definiert, deren Übergänge allerdings fließend sind:

1. Reproduzieren,
2. Zusammenhänge herstellen,
3. Verallgemeinern und reflektieren.

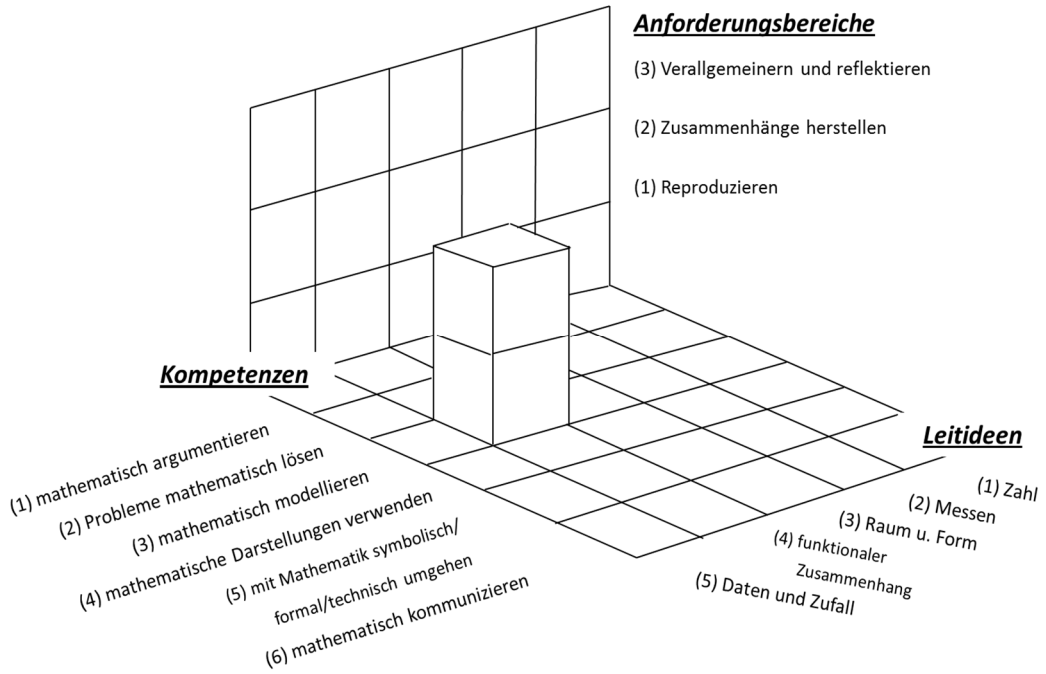


Abbildung 8: Dimensionen mathematischer Kompetenzen entsprechend der Bildungsstandards für den Hauptschul- und mittleren Bildungsabschluss im Fach Mathematik

sechs mathematische Kompetenzen

Als mathematische *Kompetenzen* werden schließlich „zentrale Aspekte mathematischen Arbeitens“ (Blum, Drüke-Noe, Hartung & Köller, 2006) differenziert. Diese Kompetenzen stehen im Zentrum der Anforderungen aus den Bildungsstandards und beschreiben pragmatisch formuliert, welche Aktivitäten Schülerinnen und Schüler beim Bearbeiten von Mathematikaufgaben beherrschen sollen. Obwohl sich die Kompetenzen am Unterricht orientieren, stellen sie doch prinzipielle mathematische Arbeitsweisen dar, die auch außerhalb der Schule das „Mathematik betreiben“ charakterisieren. Die sechs mathematischen Kompetenzen (K) beinhalten:

Mathematisch Argumentieren: Mathematisches Argumentieren beinhaltet die Formulierung logischer Argumentationsketten sowie deren Verständnis und Bewertung. Im Zentrum steht die Rechtfertigung von Behauptungen auf Basis fundamentaler mathematischer Gesetze und Konventionen, wobei diese Rechtfertigung von einfachen Plausibilitätsüberlegungen bis zu strengen Beweisen reichen kann.

Probleme mathematisch lösen: Mathematische Probleme stellen Aufgaben dar, bei denen ein unmittelbarer Lösungsweg nicht zu erkennen ist und stattdessen unter Nutzung entsprechender Strategien (z.B. heuristischer Prinzipien oder Hilfsmittel) entwickelt werden muss. Darüber hinaus ist auch die Reflexion über entsprechende Strategien Teil dieser Kompetenz.

Mathematisch modellieren: Beim Modellieren steht die vereinfachte Abbildung der Realität, entweder realer Phänomene („deskriptive Modelle“) oder realer Sachverhalte („normative Modelle“), im Mittelpunkt. Ziel ist es, Sachverhalte einer Bearbeitung zugänglich zu machen, was Verständnis und Strukturierung der Realität voraussetzt und die Anforderung beinhaltet, eine Verbindung zwischen außermathematischem Kontext und innermathematischem Inhalt zu schaffen.

Mathematische Darstellungen verwenden: In diesem Kompetenzbereich werden Aspekte zusammengefasst, bei der die eigenständige Erzeugung und Veränderung sowie der Umgang mit und das Verständnis

von mathematischen Darstellungen im Mittelpunkt stehen. Mathematische Darstellungen sind dadurch gekennzeichnet, dass diese Darstellungen (Abbildungen, Diagramme etc.) explizit mathematische Informationen beinhalten und nicht bloß der Illustration oder Motivation dienen.

Mit Mathematik symbolisch, formal und technisch umgehen: Der fünfte Kompetenzbereich beinhaltet sowohl den Gebrauch mathematischer Fakten („Wissen, dass …") als auch mathematischer Fertigkeiten („Wissen, wie …"). Dies bezieht sich auf mathematische Definitionen, Regeln, Algorithmen oder Formeln, welche bekannt sein und angewendet werden sollen. Darüber hinaus wird hier auch formales Arbeiten mit Variablen, Termen, Gleichungen oder Funktionen und die Einhaltung einer bestimmten Schrittfolge bei der Aufgabenlösung erfasst. Auch der Umgang mit Hilfsmitteln, wie Formelsammlung oder Taschenrechner, wird berücksichtigt.

Mathematisch kommunizieren: In diesem Kompetenzbereich wird das Verstehen von mathematischen Texten oder mündlichen Beiträgen zur Mathematik berücksichtigt, außerdem verständliches schriftliches oder mündliches Präsentieren mathematischer Inhalte. Im Gegensatz zum Argumentieren lässt das Kommunizieren einen externen Adressatenbezug erwarten, sodass sprachliche Erläuterungen ins Gewicht fallen.

Aufgabenformate im Fach Mathematik bei VERA 8

Für die Vergleichsarbeiten wird den Ländern derzeit eine repräsentative Aufgabenmenge zu allen Kompetenzen und Leitideen bereitgestellt. Die Aufgaben setzen sich aus einem Stimulus (Text oder Abbildung) und mehreren Items zusammen, die weitestgehend unabhängig voneinander lösbar sind. Als Aufgabenformate kommen Multiple Choice, Ergänzungsaufgaben oder andere Formate mit Kurzantworten in Frage. Teilweise wird die ausführliche Darlegung des Lösungswegs erwartet. Ungewöhnlich für viele Mathematiklehrerinnen und -lehrer ist das an Instrumenten des Bildungsmonitoring angelehnte Auswertungsschema der Aufgaben, welches nur die Optionen „Aufgabe gelöst"/„Aufgabe nicht gelöst" vorsieht (vgl. S. 67 oben).

2.4.3 Weiterführende Literatur

Literatur

Informationen über die Vergleichsarbeiten in der dritten Jahrgangsstufe sind den Internetseiten des IQB (http://www.iqb.hu-berlin.de/) und der VERA-Projektgruppe (http://www.uni-landau.de/vera/) sowie den Internetseiten der jeweiligen Ministerien und Landesinstitute der Länder zu entnehmen. Bremerich-Vos, Granzer, Behrens & Köller (2009) liefern eine Beschreibung der Bildungsstandards in der Primarstufe für das Fach Deutsch. Das Analogon für das Fach Mathematik stammt von Walther, van den Heuvel-Panhuizen, Granzer & Köller (2007).

Die aktuellen Informationen über die Inhalte der Vergleichsarbeiten in der achten Jahrgangsstufe sind ebenfalls auf den Internetseiten des IQB bzw. den Webseiten der Ministerien und Landesinstitute zu finden. Eine detaillierte Beschreibung der Kompetenzbereiche in den Bildungsstandards in der Sekundarstufe I ist von Bremerich-Vos, Granzer und Köller (2008) vorgelegt worden. Informationen zur Normierung der Bildungsstandards in Englisch sind dem technical report[10] des IQB zu entnehmen (Rupp, Vock, Harsch & Köller, 2008). Konzept und erste Ergebnisse zur Normierung in Französisch – hier wie gesagt nicht vorgestellt – stellen Tesch, Leupold & Köller (2008) dar. Das Kompetenzstufenmodell und weiterführende Informationen zu den Bildungsstandards im Fach Mathematik erläutern Blum, Drüke-Noe, Hartung & Köller (2006).

Die Bildungsstandards für die Primarstufe, den Haupt- und mittleren Bildungsgang sind online auf den Webseiten der KMK (http://www.kmk.org/bildung-schule/qualitaetssicherung-in-schulen/bildungsstandards/ueber-blick.html) oder im Kluwer Verlag verfügbar. Zusätzliche Informationen zur Beschreibung der jeweiligen Kompetenzstufenmodelle sind beim IQB verfügbar (http://www.iqb.hu-berlin.de/bista?reg=r_4). Als Ergänzung empfiehlt sich die Expertise zur Entwicklung der Standards von Klieme et al. (2003). Ein lesenswertes Plädoyer für die Qualitätssicherung schulischer Arbeit auf der Basis von Bildungsstandards hält Köller (2008).

2.4.4 Verständnis und Diskussionspunkte

1. *Diskutieren Sie, welche Vorteile sich aus einer (bundesweit) zentral vorgenommenen Testentwicklung für Vergleichsarbeiten ergeben. Welche Schwierigkeiten könnten gleichzeitig damit verbunden sein?*
2. *Die Wahl geschlossener Antwortformate, wie etwa Multiple Choice, für die Erfassung von Lese- und Hörverstehen wird auch mit der Validität der Tests begründet. Wie ist dies zu verstehen?*

10 Der technical report ist als Online-Publikation verfügbar unter: http://www.iqb.hu-berlin.de/arbbereiche/projekte?pg=p_7&spg=r_8

3. *Im Fach Englisch sind die Testaufgaben an den GER angebunden. Welche Vor- und Nachteile sind mit der Verwendung des GER in diesem Zusammenhang verbunden?*
4. *Vom IQB wird an verschiedenen Stellen betont, dass Testaufgaben keine optimalen Lernaufgaben darstellen. Worin könnten grundlegende Unterschiede bestehen?*

2.5 Praktische Implikationen: Ergebnisnutzung aus Vergleichsarbeiten und Unterrichtsentwicklung

Im letzten Kapitel dieses Studienbriefs werden die unmittelbaren Implikationen von Vergleichsarbeiten für die schulische Tätigkeit vorgestellt. Es wird zunächst das Format der Rückmeldungen an einigen Beispielen erläutert. Im Anschluss wird ein Vorschlag zur Umsetzung der aus Vergleichsarbeiten gewonnenen diagnostischen Informationen für pädagogische und didaktische Entwicklungsprozesse (als vielleicht wichtigster Zielsetzung) gemacht. Danach folgt die Beschreibung eines Schulbeispiels, an dem aufgezeigt wird, welchen Einfluss Vergleichsarbeiten auf die Schulpraxis haben können. Abschließend wird die Brücke zum kompetenzorientierten Unterricht geschlagen.

In diesem Kapitel werden folgende Fragen beantwortet:

- *Wie lassen sich die Ergebnisrückmeldungen aus Vergleichsarbeiten interpretieren?*
- *Welche Evaluationsschritte sind im Anschluss an die Ergebnisinterpretation zu durchlaufen, damit Unterrichtsentwicklungsprozesse in Gang gesetzt werden können?*
- *Wie gelangt man von den Ergebnisrückmeldungen der Vergleichsarbeiten zu einem kompetenzorientierten Unterricht?*

2.5.1 Interpretation der Rückmeldungen aus Vergleichsarbeiten

Ergebnisinterpretation

Mit der Ergebnisinterpretation aus Vergleichsarbeiten ist das Ziel verbunden, unter der Nutzung kriterialer und sozialer Vergleiche Prozesse der Schul- und Unterrichtsentwicklung in Gang zu setzen. Lehrerinnen und Lehrern der an den Vergleichsarbeiten beteiligten Klassen wird umfangreiches Begleitmaterial zur Verfügung gestellt, welches unter anderem Anregungen zur Nutzung der Testergebnisse und -materialien im Unterricht beinhaltet. Nahezu alle Materialien der Vergleichsarbeiten können in der Regel, wie die Ergebnisse der Schule, über passwortgeschützte Internetportale abgerufen werden, die die Länder zu diesem Zweck eingerichtet haben. Darstellung und Umfang der Rückmeldungen variieren zwischen den Bundesländern. Zumeist umfassen sie aber:

Materialien

- Testaufgaben und Auswertungsmanuale,
- Lösungshäufigkeiten der einzelnen Aufgaben,
- Kommentierungen der Aufgaben und Hinweise für die Weiterarbeit mit den Aufgaben,
- eine Ergebnisdarstellung der eigenen Klassen und Jahrgangsstufen (z.B. als Verteilung von Schülerinnen und Schülern auf Kompetenzstufen),
- kriteriale Beschreibung der Kompetenzanforderungen,
- Angabe von Referenzwerten (beispielsweise Ergebnisse von Schulen, welche unter vergleichbaren Rahmenbedingungen arbeiten) sowie (statistische) Interpretationshilfen für einen sozialen Vergleich.

Diagnostisches Potential von Vergleichsarbeiten

Das diagnostische Potential der Testaufgaben liegt darin, differenziert Stärken und Schwächen der Schülerschaft in Bezug auf einzelne Inhalte zu untersuchen (vgl. Möller, Pallack & Fleischer, 2007; Kliemann, 2010). Dennoch soll im Folgenden die kriteriale und soziale Interpretation im Hinblick auf *Kompetenzausprägungen* im Zentrum der Darstellung von Ergebnisrückmeldungen stehen. Um die Darstellung nicht zu abstrakt werden zu lassen, wird wiederum auf konkrete Beispiele der Rückmeldungen aus den vergangenen Jahren zurückgegriffen.

Beispiel:

Abbildung 9 zeigt eine Schulrückmeldung des Kompetenzbereichs Leseverstehen im Unterrichtsfach Deutsch, Kompetenzbereich Leseverstehen, der Lernstandserhebungen in Nordrhein-Westfalen im Jahre 2011. Die Rückmeldung beinhaltet für eine ausgewählte Klasse (oben), die Jahrgangsstufe (Mitte) und die Parallelklassen (unten) der Beispielschule die prozentuale Verteilung der Schülerinnen und Schüler auf fünf Kompetenzstufen zuzüglich einer Kategorie „k.h.N.“. Liegen für eine Schülerin/einen Schüler keine oder unvollständige Daten vor, ist eine Zuordnung zu den beschriebenen Kompetenzniveaus nicht möglich; dies wird als „kein hinreichender Nachweis für das Erreichen eines Kompetenzniveaus” bezeichnet, d. h. in diesem Bereich wurden in der Regel nur vereinzelt Aufgaben gelöst. Die ins-

gesamt geringe Anzahl dieser gelösten Aufgaben ermöglicht es jedoch nicht, die Schülerin bzw. den Schüler mit hinreichender Sicherheit einem Kompetenzniveau zuzuordnen. Der rechte Rand der Abbildung zeigt Erläuterungen zu den Kompetenzstufen. Lehrerinnen und Lehrer werden hier zu den Kompetenzanforderungen der jeweiligen Stufe weitergeleitet.

Der Abbildung ist zunächst im Sinne eines kriterialen Vergleichs zu entnehmen, dass der überwiegende Teil der ausgewählten Klasse (73 %) die Anforderungen der Kompetenzstufen 3, 4 und 5 nicht erfüllt. Lediglich 17% Schülerinnen und Schüler erreichen die Stufe 3, 9 % der Schülerinnen und Schüler erzielen Leistungen auf Stufe 4. Ein ähnliches Ergebnis findet sich auch für die gesamte Jahrgangsstufe (Mitte), in der 78% der Schülerinnen und Schüler nicht die Kompetenzstufe 3 erreichen. Für eine möglichst objektive Einschätzung dieser Ergebnisse ist es bedeutsam, mit welcher Referenzgruppe sich die Schule vergleicht. Die Schule ist auf Basis des in Kapitel 2.2.2.1 vorgestellten Verfahrens der Referenzgruppe einem sogenannten Standorttyp, dem Standorttyp 5 mit den ungünstigsten Eingangsvoraussetzungen der Schülerschaft, zugeordnet worden. Diese Referenzgruppe ermöglicht einen Vergleich mit anderen Schulen, an denen unter ähnlichen Rahmenbedingungen gearbeitet wird.

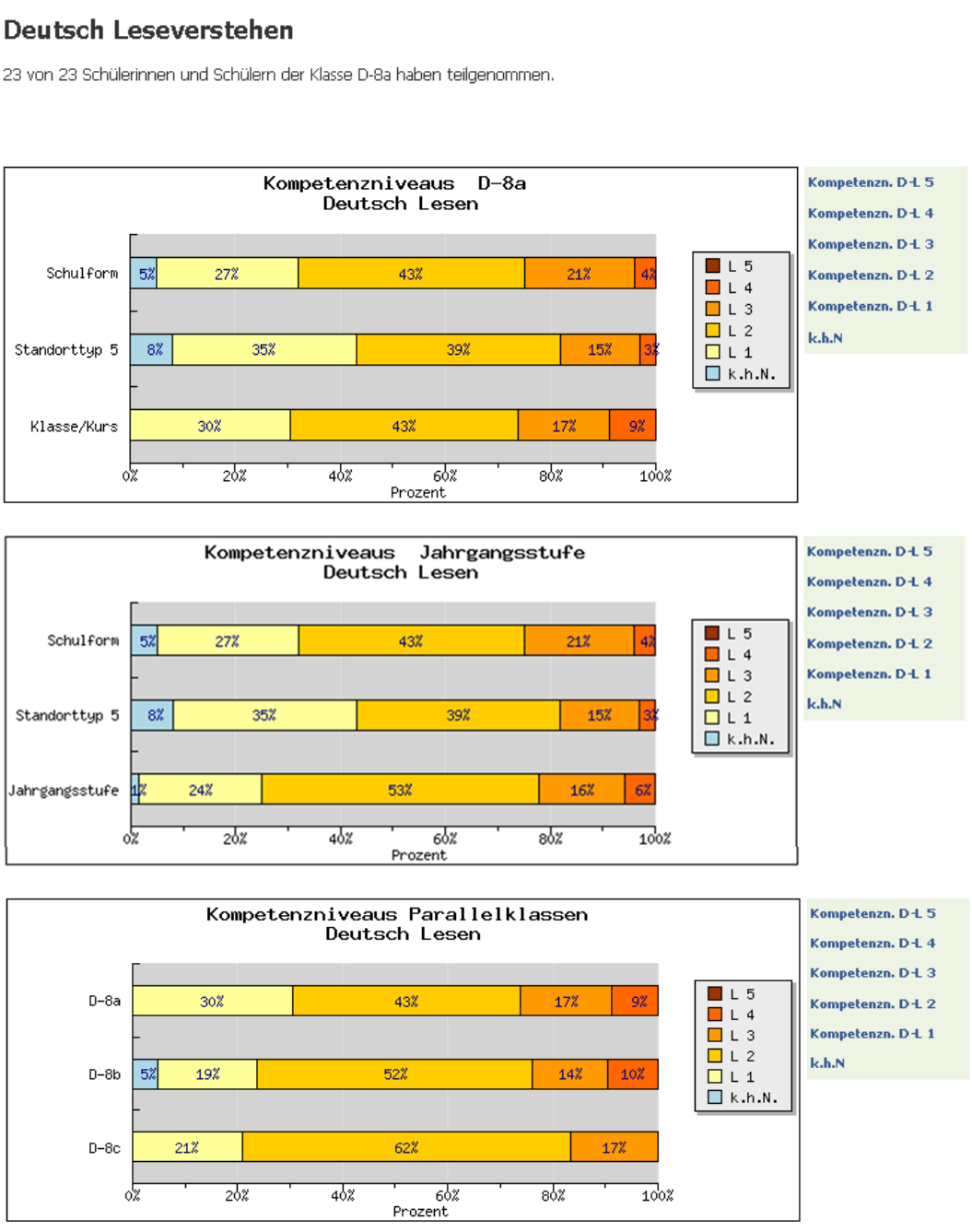


Abbildung 9: Ergebnisrückmeldung Lernstand 8 in Nordrhein-Westfalen des Jahres 2011 (Beispiel für eine Klasse im Fach Deutsch, Kompetenzbereich Leseverstehen)

Die Berücksichtigung der Referenzverteilungen von Standorttyp 5 im Sinne eines sozialen Vergleichs relativiert das Abschneiden der Klasse und Jahrgangsstufe, denn in diesem Standorttyp verfehlen insgesamt sogar 82% das Kompetenzniveau 3. Um zu verhindern, dass marginale und eher zufällige Unterschiede zwischen der eigenen und den Referenzverteilungen interpretiert werden, erhalten Lehrkräfte zu den Abbildungen jeweils kurze Hinweistexte.

Der Vergleich mit den Parallelklassen (unten) liefert viele Hinweise für die innerschulische Diskussion, da innerhalb der Schule detaillierte Informationen zu dem in den Parallelklassen realisierten unterrichtlichen Vorgehen verfügbar sind. Bei der vorgestellten Schule fällt beispielsweise auf, dass die Klasse 8c homogener abschneidet als ihre Parallelklassen, deren Ergebnisse stärker streuen.

In Kapitel 2.2 war bereits darauf eingegangen worden, dass bei den Kompetenztests in Thüringen ein anderer Weg der Ergebnisrückmeldung gewählt worden ist. Die Klassen erhielten bei den Kompetenztests die durchschnittlich erreichte absolute und prozentuale Punktzahl zurückgemeldet. Diese Punktzahlen konnten mit einem auf Basis schulleistungsrelevanter Hintergrundmerkmale korrigierten Landesmittelwert (siehe Kapitel 2.2) verglichen werden. Abbildung 10 zeigt eine Ergebnisrückmeldung des thüringischen Kompetenztests in der dritten Jahrgangsstufe aus dem Jahr 2009.

Die Abbildung führt auf der linken Seite in Form eines Balkendiagrammes die durchschnittliche prozentuale Punktzahl im Test für die Beispielklasse (rot) und den korrigierten Landesmittelwert (gelb) auf. Dieser korrigierte Landesmittelwert ist aufgrund sozioökonomischer Hintergrundmerkmale der Schülerschaft in der Beispielklasse 3Z berechnet worden und ermöglicht einen „fairen“ sozialen Vergleich der Schulleistungen in dieser Klasse (vgl. Kapitel 2.2). Auf der rechten Seite erhält die Lehrkraft zusätzlich einige deskriptive Statistiken[11] der erzielten Ergebnisse. Sie kann sich darüber beispielsweise einen Eindruck von der Leistungsstreuung in ihrer Klasse (minimal erzielte Punktzahl/ maximal erzielte Punktzahl) machen und erfährt die maximal erreichbaren Punkte sowie die absolute und relative Lösungsquote der eigenen Klasse und des korrigierten Landesmittelwerts.

Dem Balkendiagramm ist zu entnehmen, dass in der Beispielklasse fast 80 % der Aufgaben gelöst wurden und sie damit im Vergleich mit dem (korrigierten) Landesdurchschnitt (72 %) besser abschneidet. Den deskriptiven Statistiken lässt sich des Weiteren entnehmen, dass der Unterschied zwischen der Klasse und dem (korrigierten) Landesdurchschnitt bei 2,5 Punkten liegt. Die Streuung in der Beispielklasse umfasst 24 Punkte (59 %) und erscheint damit auf den ersten Blick recht groß. So erzielt die schwächste Schülerin bzw. der schwächste Schüler der Klasse mit 16 gelösten Items gut 40 % der möglichen Punkte, die beste Schülerin oder der beste Schüler verfehlt mit 40 Punkten (98 %) nur knapp die Höchstpunktzahl.

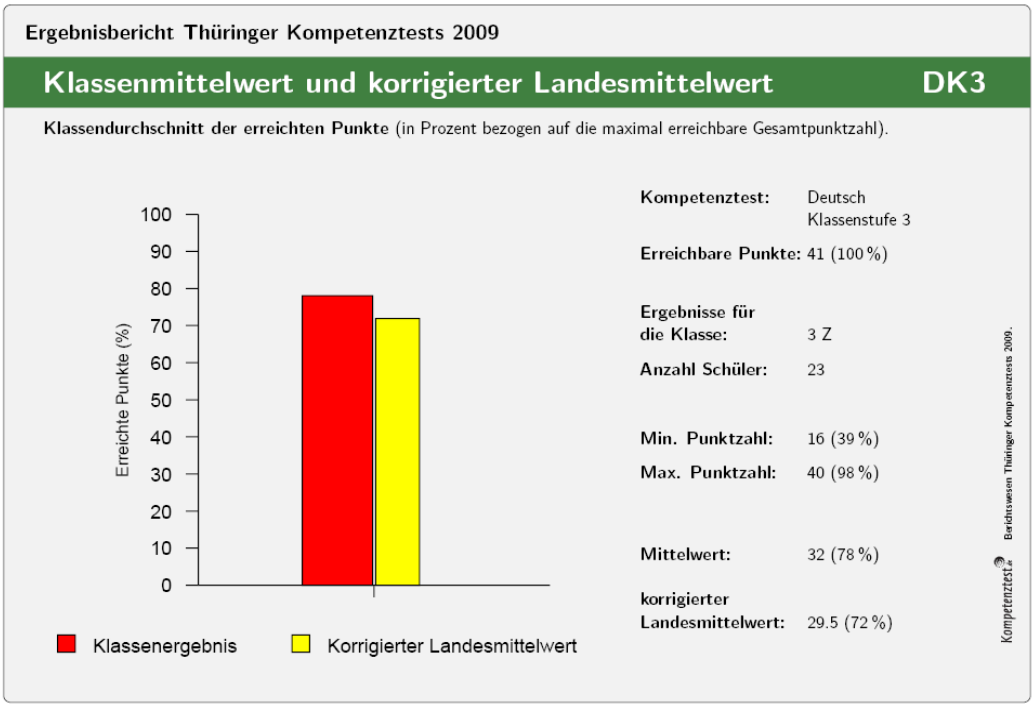


Abbildung 10: Ergebnisrückmeldung der „Kompetenztest “ in Thüringen 2009 (Beispiel für eine Klasse im Fach Deutsch)

11 Deskriptive Statistiken fassen die Ergebnisse der Klasse anhand geeigneter statistischer Maße (z.B. Mittelwert und Standardabweichung) zusammen.

Weitere Informationen über die Leistungsstreuung innerhalb der Klasse sind der Abbildung 11 zu entnehmen, die sich ebenfalls auf die Beispielklasse 3Z bezieht. Die Abbildung zeigt einen so genannten Boxplot der Leistungsstreuung in der Beispielklasse (hervorgehoben oben) und dem Landesmittelwert (unten). Das mittlere graue Segment des Boxplots umfasst die mittleren 50 % der Schülerinnen und Schüler in der Klasse bzw. des Landes, wobei durch einen senkrechten Strich der Median angegeben ist. Der Median beschreibt jenen Testwert, unter und über dem 50 % der getesteten Schülerinnen und Schüler liegen; er halbiert also die Verteilung. Links des grauen Segmentes und der Markierung „25 %“ liegt das Viertel der leistungsschwächsten Schülerinnen und Schüler. Rechts des grauen Segmentes und der Markierung „75 %“ liegt das Viertel der leistungsstärksten Schülerinnen und Schüler. Das Ende des roten bzw. blauen Bereiches markiert den Beginn der extremsten 10 % der Klasse bzw. des Landes, die dahinter liegenden „Whiskers“ (Schnurrbarthaare) geben in der Verteilung der Klasse die jeweils leistungsschwächste (Minimum) bzw. leistungsstärkste (Maximum) Person an. Die Abbildung verdeutlicht, dass im Landesdurchschnitt die Streuung der Testergebnisse in der unteren Leistungshälfte (rechts des Median) größer ist als in der oberen Leistungshälfte. Insbesondere das leistungsstärkste Quartil des Landes ist in seinen Leistungen vergleichsweise homogen, was allerdings nicht überrascht, da die Lösungsquote des Tests relativ hoch ist. Dieses Ergebnis spiegelt sich auch beim Blick auf den Boxplot der Beispielklasse wider. Revidiert werden muss hingegen der erste Eindruck von der Leistungsstreuung in Klasse 3Z, denn der Abstand zwischen den stärksten und schwächsten 10 % der Beispielklasse ist geringer als im Landesdurchschnitt. Der Lehrkraft in der Beispielklasse ist es also offenbar gelungen, die Leistungsstärke in der Klasse zu homogenisieren. Es kann an diesem Beispiel verdeutlicht werden, wie bedeutsam die Berücksichtigung von Referenzwerten ist: Wird der soziale Vergleich ignoriert, können Fehlinterpretationen entstehen. Der Abbildung ist zudem zu entnehmen, dass die Beispielklasse insbesondere im unteren Leistungsbereich (rotes Segment des Boxplots) homogener abschneidet als der Landesdurchschnitt. Insbesondere recht schwache Leistungen sind also in der Beispielklasse selten. Im Median unterscheiden sich beide Boxplots dagegen kaum. Dies bestätigt den Eindruck, dass das bessere durchschnittliche Abschneiden der Beispielklasse im Wesentlichen auf die geringere Streuung im unteren Leistungsbereich zurückzuführen ist.

Die Beispiele verdeutlichen, dass für eine verständnisvolle Nutzung der Ergebnisrückmeldungen eine Auseinandersetzung mit dem jeweiligen Rückmeldeformat notwendig ist. So lassen sich aus den Abbildungen und den zusätzlichen Informationen der beiden Beispiele unterschiedliche Erkenntnisse über das Abschneiden der eigenen Klasse (und ggf. der eigenen Jahrgangsstufe) ziehen. Statistische Grundkenntnisse werden dabei vorausgesetzt. Lehrkräften, denen diese fehlen, kann als Literatur z. B. Eikenbusch und Leuders (2004) oder Lienert und von Eye (1998) empfohlen werden.

Die Rückmeldeformate der Bundesländer variieren, sodass die Formate in einigen Ländern von den hier dargestellten Beispielen abweichen können. In der Regel stellen die Länder aber auf ihren Webseiten Beispiele für die Ergebnisinterpretation der landespezifischen Formate bereit.

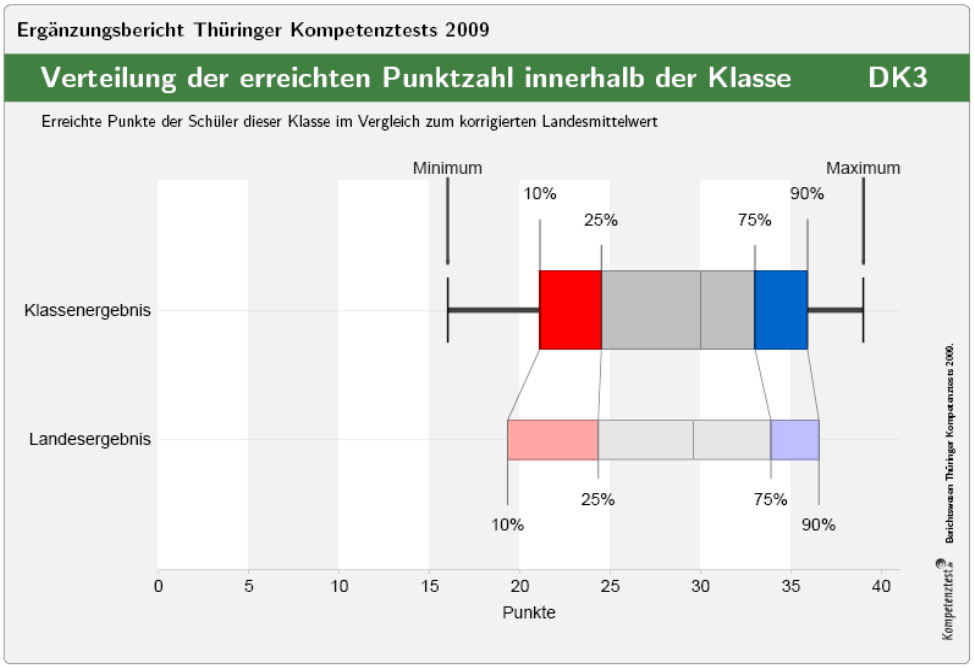


Abbildung 11: Ergänzende Informationen zur Ergebnisrückmeldung der „Kompetenztests “ in Thüringen 2009 (Beispiel für eine Klasse im Fach Deutsch)

2.5.2 Von der Ergebnisinterpretation zur schulischen Entwicklung

Wie kann die zentrale Zielsetzung der Unterrichtsentwicklung auf Basis der Ergebnisinterpretation aus Vergleichsarbeiten realisiert werden? Im Folgenden werden vier zentrale Schritte zur Nutzung der Ergebnisrückmeldungen für eine schulinterne Evaluation vorgestellt. Sie sind grob am Zyklenmodell von der Evaluation zur Innovation nach Helmke und Hosenfeld (2004) orientiert, werden aber für die entsprechende Zielsetzung hier konkretisiert.[12]

1) Ergebnisinterpretation anhand kriterialer und sozialer Bewertungsmaßstäbe:

Ergebnis-interpretation

Eine zutreffende Interpretation ist Voraussetzung für die verständnisvolle Ergebnisnutzung. Die kriteriale Interpretation einer Ergebnisrückmeldung beinhaltet beispielsweise folgende Fragen (vgl. auch Peek & Dobbelstein, 2006):

kriteriale Interpretation

Wie viele meiner Schüler erreichen welche Kompetenzerwartungen?
Lassen sich anhand der Fehlermuster typische Kompetenzlücken identifizieren?

soziale Vergleiche

Relevant wäre z.B., ob Fehler systematisch in bestimmten **Inhaltsbereichen** oder bei bestimmten **Aufgabentypen** auftreten. Vor dem Hintergrund sozialer Vergleiche kann eine Lehrkraft untersuchen:

Wie schneidet meine Klasse innerhalb der Jahrgangsstufe ab? Bestehen Leistungsunterschiede zwischen meiner Klasse und den Parallelklassen? Sind Stärken und Schwächen in bestimmten Bereichen identifizierbar (differenzielle Leistungsunterschiede)? Wie schneidet meine Klasse im Vergleich mit Klassen anderer Schulen mit ähnlicher Schülerzusammensetzung ab? Finden sich Hinweise, dass meine Lerngruppe entsprechend dieser Referenz ihre Möglichkeiten nicht ausschöpft?

Zudem bietet es sich an, die Ergebnisse zu Zeugnisnoten, Klassenarbeiten und Parallelarbeiten ins Verhältnis zu setzen. Dabei sollte das Gesamtbild der Klasse betrachtet werden und bzgl. systematischer Abweichungen geprüft werden.

2) Reflexion und Ursachenanalyse in den Fachkonferenzen:

Individuelle Reflexion und Diskussion der Ergebnisse

Vergleichsarbeiten sollen Lehrkräfte in die Lage versetzen, in den Fachkonferenzen über die Ursachen für die festgestellten Ergebnisse in Diskussion zu kommen und Anhaltspunkte für mögliche Interventionen herauszuarbeiten. Es sollten insbesondere berücksichtigt werden:

- Besonderheiten der Schülerschaft
- inhaltliche Überschneidungen mit anderen Unterrichtsfächern, z.B. durch bilingualen Unterricht oder Überschneidungen zwischen Mathematik- und Informatikunterricht

3) Folgerungen aus den Ergebnissen und Interventionen:

Folgerungen aus den Ergebnissen / Interventionen

Im Sinne des Wechsels von einer Input- zu einer Outputorientierung im Bildungswesen liegt es im Verantwortungsbereich der Schule, unter Berücksichtigung der eigenen schulischen Besonderheiten selbstständig geeignete Konsequenzen aus den Ergebnisrückmeldungen zu entwickeln. Es sollen hier nur grob einige Ansatzpunkte beschrieben werden:

- Gemeinsame Unterrichtsvorbereitung oder Hospitationen in Parallelklassen können den eigenen Referenzrahmen für die zu erwartenden Leistungen von Schülerinnen und Schülern erweitern und die eigene diagnostische (Bewertungs-)Kompetenz stärken. Unterschiedliche Unterrichtsmaterialien und Leistungsanforderungen der Fachlehrerinnen und -lehrer können zudem einander gegenübergestellt werden.
- Die Passung zwischen schulinternem Curriculum und den in den Bildungsstandards skizzierten Kompetenzanforderungen sollte untersucht werden. Didaktische Trainer können sinnvolle Unterstützung bieten, um neue Inhalte in das schulinterne Curriculum zu integrieren.
- Individuelle Fördermaßnahmen, insbesondere für die Gruppe der leistungsschwachen Schülerinnen und Schüler, gewinnen oftmals an Bedeutung. In diesem Fall ist zu prüfen, ob sich die Ergebnisse der Vergleichsarbeiten in ein konsistentes Bild über diese Schülerinnen und Schüler einfügen lassen, was eine systematische Dokumentation der Leistungsentwicklung voraussetzt. Im Anschluss können geeignete Fördermaßnahmen (Förderstunden, Binnendifferenzierung, etc.) in Erwägung gezogen werden.

4) Schulinterne Evaluation der Interventionsmaßnahmen:

Evaluation von Interventions-maßnahmen

Hat eine Intervention überhaupt Wirkung gezeigt und wie nachhaltig ist die Wirkung? Um dies beurteilen zu können, bietet es sich zum Beispiel an, zu einem späteren Zeitpunkt dieselben oder parallele Item-Stichproben der Vergleichsarbeiten erneut zu verwenden. Es kann so anhand repräsentativer Lösungs-

12 Neben dem hier beschriebenen Rezeptionsverlauf beinhaltet das Modell individuelle und externe Rahmenbedingungen, die hier nicht angeführt werden.

quoten erfasst werden, wie groß der Lernfortschritt von Schülerinnen und Schülern in einem bestimmten Zeitintervall ist und ob sich Profile von Kompetenzstärken und -schwächen verändert haben. Lehrerinnen und Lehrern werden künftig verstärkt Materialien zur Lernstandsdiagnose bereitgestellt. Das IQB plant, den Schulen zum Zweck der Selbstevaluation den Abruf von identischen oder parallelen Testaufgaben der Vergleichsarbeiten in der achten Jahrgangsstufe aus einer Itemdatenbank zu ermöglichen. Ein ähnliches Vorhaben verfolgen in der Mathematik auch die Projekte „SINUS-Transfer" (Sekundarstufe I; vgl. auch Studienbrief 3) und „SINUS Transfer Grundschule“, welche ebenfalls Materialien zur regelmäßigen Lernstandsdiagnose zur Verfügung stellen. Lehrerinnen und Lehrer haben so die Möglichkeit, zu einem ihnen günstig erscheinenden Zeitpunkt die Entwicklung des Kompetenzerwerbs ihrer Schülerinnen und Schüler festzustellen.

Mit den hier vorgestellten vier Schritten der internen Evaluation ist der schulische Entwicklungsprozess natürlich nicht abgeschlossen. Die Evaluationsbefunde gehen vielmehr in die Interpretation weiterer diagnostischer Informationen ein, sodass sich ein Zyklus regelmäßiger Evaluationsmaßnahmen ergibt. Abbildung 12 fasst die vier vorgestellten Schritte noch einmal zusammen.

2.5.3 Ein Beispiel für die Ergebnisnutzung in den Fachkonferenzen

Fachkonferenzen dürfen in den Schulen als ein zentrales Gremium der Qualitätsentwicklung betrachtet werden. In ihrem Aufgabenbereich liegen (i.d.R.) zahlreiche didaktische und methodische Arbeiten, u.a.:

- die Ausarbeitung des schulinternen Curriculums und gemeinsamer Unterrichtsreihen
- die Evaluation geeigneter Unterrichtsmaterialien
- die Organisation von Hospitationen und Weiterbildungen.

Wie bereits dargelegt (vgl. den vorangegangenen Abschnitt), sind Fachkonferenzen mit diesem Verantwortungsbereich auch geeignete Adressaten der Rückmeldungen aus Vergleichsarbeiten. Zur Verdeutlichung, wie die Ergebnisse der Vergleichsarbeiten in den Fachkonferenzen diagnostisch genutzt werden können, wird im Folgenden die Arbeit einer Fachkonferenz Deutsch im Anschluss an eine Ergebnisrückmeldung zum Kompetenzbereich „Leseverstehen“ als Beispiel dargestellt.

In der Auseinandersetzung mit den Ergebnissen hatte die Fachkonferenz auf Basis von Vorarbeiten der Jahrgangsstufenlehrerinnen und -lehrer beschlossen, folgende Punkte zu thematisieren:

- Allgemein: Welche Ergebnisse bedürfen einer besonderen Beachtung?
- Wie gestaltet sich der Parallelklassenvergleich? Lassen sich aus den Ergebnissen Informationen gewinnen, welche in den schulinternen Vergleichen (bspw. Parallelarbeiten) nicht offensichtlich geworden sind?
- Wie gestaltet sich der Vergleich der Jahrgangsstufe mit Referenzschulen?
- Ergeben sich systematische Unterschiede zwischen Schülergruppen, bspw. zwischen Mädchen und Jungen?
- Welche Konsequenzen ergeben sich aus den Ergebnissen für die vorausgegangenen bzw. die nachfolgenden Jahrgangsstufen? Wie lässt sich in diesen Jahrgangsstufen kompetenzorientierter Unterricht erfolgreich implementieren bzw. weiterentwickeln?

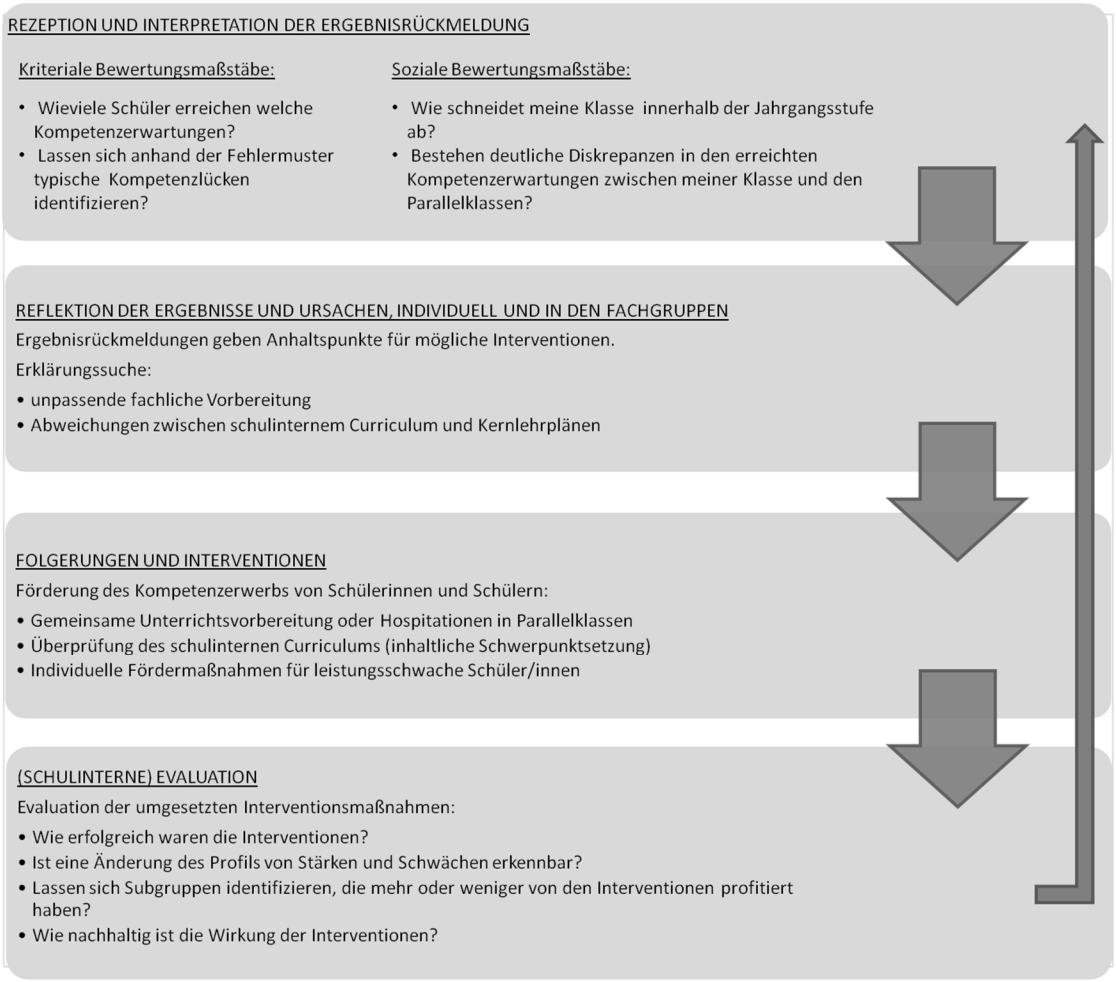


Abbildung 12: Schema für das Vorgehen von der Interpretation der Ergebnisrückmeldung bis zur Ergebnisnutzung für die schulinterne Evaluation

Zwei der hier aufgeführten Diskussionspunkte seien exemplarisch herausgegriffen, um konkret darzustellen, welche Konsequenzen die Fachkonferenz aus den Ergebnissen gezogen hat:

Der Parallelklassenvergleich der Ergebnisse zeigte deutliche Unterschiede zwischen den Klassen. So schnitten zwei der vier Parallelklassen signifikant besser ab als entsprechende Referenzgruppen, wohingegen die anderen beiden Klassen der Gesamtschule im Durchschnitt blieben. Die Fachkonferenz verglich die erzielten Ergebnisse mit früheren Leistungsvergleichen der Klassen, welchen derartige Leistungsunterschiede jedoch nicht zu entnehmen waren. Diskutiert wurden Unterschiede zwischen den Klassen in der Zusammensetzung der Schülerschaft (Lernmotivation der Klasse, Anteil der Deutsch-Muttersprachler, Mädchenanteil; vgl. den nachfolgenden Absatz). Die Mitglieder zeigten sich jedoch skeptisch, ob Unterschiede in der Schülerzusammensetzung die Ergebnisse allein erklären könnten. Zur Sprache kam im Anschluss auch die Frage, inwiefern eine unterschiedliche Bewertungsstrenge der Fachlehrerinnen und -lehrer eine Rolle spielen könnte. Als weitreichende Konsequenz schlug die Fachkonferenz der Schulleitung die Einführung regelmäßiger Parallelarbeiten vor, um einen besseren schulinternen Vergleich zu ermöglichen und innerhalb der Fachkonferenz einen noch intensiveren Austausch über Unterrichtsinhalte, Aufgaben und Bewertungskriterien anzuregen. Nach anfänglicher Skepsis werden diese in den Kernfächern inzwischen in allen Jahrgangsstufen der Sekundarstufe I einmal pro Schuljahr durchgeführt und stehen im Kollegium inzwischen als hilfreiches Instrument des sozialen Vergleichs außer Zweifel. Der Nutzen von Vergleichsarbeiten für einen darüber hinaus gehenden externen Vergleich mit anderen Schulen ist ebenfalls richtig wahrgenommen worden.

Ein weiteres interessantes Ergebnis der Vergleichsarbeiten beinhaltete Geschlechtsunterschiede in der Lesekompetenz: Den Ergebnisrückmeldungen der Gesamtschule war deutlich zu entnehmen, dass Jungen im Durchschnitt schwächer abschnitten als Mädchen. Die Fachkonferenz bezog in die Diskussion der Ergebnisse der Vergleichsarbeiten auch Ergebnisse eines standardisiert durchgeführten Sprachstandtests zu Beginn der fünften Jahrgangsstufe ein, die im Lesen ebenfalls einen deutlichen Geschlechtsunterschied zugunsten der Mädchen feststellen ließen. Offenbar wurden schwächere Leseleistungen der

Jungen also aus der Grundschule „verschleppt“. Ausgehend von der Feststellung, dass schwächere Leseleistungen nicht zuletzt auch aus einem geringeren Leseinteresse und einer niedrigeren Lesefrequenz erklärt werden können, wurde die pragmatische Entscheidung getroffen, dass bis zu drei Jungen aus jeder Klasse in der 5. und 6. Jahrgangstufe eine zusätzliche Stunde Leseunterricht (Viellese-Verfahren) erhalten sollten. Die Schüler wählen sich dazu jeweils eigenständig Bücher und trainieren, begleitet von einem männlichen Lehrer, Lesefertigkeiten und Lesestrategien. Die Fachkonferenz entschied, diese Maßnahme zu evaluieren, indem die Leseleistungen von Jungen und Mädchen nun regelmäßig in allen Jahrgangsstufen verglichen werden. Der bei den Vergleichsarbeiten NRW in regelmäßigen Zyklen wiederkehrende Kompetenzbereich „Leseverstehen“ eröffnet der Fachkonferenz zusätzlich die Möglichkeit, die verwendeten Leseaufgaben zu reflektieren und Referenzwerte für die Leseleistungen von Jungen und Mädchen zu erhalten.

Im Anschluss widmete sich die Fachkonferenz der Frage, wie Methoden des kompetenzorientierten Unterrichts stärker zu implementieren seien und setzte sich dazu mit den die Ergebnisrückmeldung ergänzenden Materialien zur Weiterarbeit auseinander. Die Implementation kompetenzorientierten Unterrichts wird, veranschaulicht am Kompetenzbereich Leseverstehen, separat im folgenden Abschnitt behandelt.

Die Fachkonferenz beschloss ihre Arbeit mit der Ausarbeitung eines Ergebnisberichts an die Schulaufsicht. In ihm werden neben den Ergebnissen auch die spezifische Situation der Schule und die geplanten Interventionsmaßnahmen dargelegt.

2.5.4 Kompetenzorientierter Unterricht

Grundgedanke kompetenzorientierten Unterrichts

Vergleichsarbeiten ebnen den Weg für kompetenzorientierten Unterricht. Unter kompetenzorientiertem Unterricht wird hier nach Ziener (2006) verstanden, „Stoffe, Inhalte oder Themen im Unterricht so zu bearbeiten, dass dabei Kompetenzen, wie sie in den Bildungsstandards formuliert sind, angebahnt, eingeübt oder erworben werden können“. Nach den Vorstellungen der KMK (KMK & for.mat, ohne Jahr) verläuft die Planung kompetenzorientierten Unterrichts in drei Schritten:

1. Aneignung von Kenntnissen über die Inhalte der Bildungsstandards und Verständnis dieser Inhalte: Dies bezieht sich nicht allein auf die Kompetenzbeschreibungen, sondern beinhaltet auch die in den Standards aufgeführten Methoden und Arbeitstechniken (vgl. Kapitel 2.4).
2. Untersuchung der didaktischen Ziele dieser Kompetenzbeschreibungen: Die KMK spricht hier von der „Kompetenzexegese“, deren Inhalt nicht zuletzt die gegenstandsbezogene Konkretisierung der Kompetenzbeschreibungen ist.
3. Ableitung von Aufgabenstellungen und Erstellung einer Sequenzplanung durch die strukturierte Zusammenführung dieser Aufgabenstellungen: Die Aufgabenbeispiele der Vergleichsarbeiten geben Lehrkräften bereits einen Eindruck davon, wie kompetenzorientierte (Test-)Aufgaben aussehen können. Es liegt in der Verantwortung der Fachkonferenz, diese zu kompetenzorientierten Lernaufgaben weiterzuentwickeln und in geeignete Unterrichtsreihen einzubinden. Sie diskutiert, welche Veränderungen der Materialien (Kürzungen oder Erweiterungen) notwendig sind, und mit welchen Konsequenzen in Bezug auf Diagnose und Differenzierung der Einsatz der Materialien verbunden ist. Häufig bietet es sich an, innerhalb der Fachkonferenz Tandems von jeweils zwei Lehrkräften zu bilden, welche jeweils bestimmte Teile der Materialien und Aufgaben bzgl. ihrer Potentiale untersuchen und Erfahrungen im Umgang mit den Materialien dokumentieren.

Kompetenzzuwächse im Unterricht

Um Kompetenzzuwächse im Unterricht zu erzielen, wird von der KMK die Verknüpfung zweier Elemente vorgeschlagen (KMK & for.mat, ohne Jahr): zum einen die regelmäßige diagnostische Erfassung der Lernstände der Schülerinnen und Schüler, etwa im Rahmen von Vergleichsarbeiten. Zum anderen wird davon ausgegangen, dass Kompetenzzuwächse durch eine systematische Variation des Schwierigkeitsgrades des Unterrichtsgegenstandes bzw. der im Unterricht eingesetzten Aufgaben erzielt werden. Im KMK-Projekt for.mat (http://www.kmk-format.de) sind Analyseschemata entwickelt worden, welche es möglich machen, schwierigkeitsbestimmende Merkmale von Texten und Leseaufgaben zu identifizieren. Abbildung 13 gibt hierzu ein Beispiel für Lernaufgaben des Kompetenzbereichs Leseverstehen. Alternativ lässt sich die Schwierigkeit des Unterrichtsgegenstandes variieren, im Bereich Leseverstehen also beispielsweise die Textschwierigkeit. Beide Methoden stellen eine nicht zu unterschätzenden Herausforderung an die diagnostischen Fähigkeiten der Lehrerinnen und Lehrer dar, denn die Schwierigkeit von Unterrichtsgegenstand und Aufgaben muss von ihnen richtig eingeschätzt werden. Entsprechende Hilfsmittel sind allerdings inzwischen entwickelt worden (vgl. ebenfalls die zuvor bereits beschriebene Webseite von for.mat).

Im Folgenden werden zwei Beispiele zur Weiterarbeit im Unterricht vorgestellt.

Beispiel 1: Möglichkeiten zur Förderung der Lesekompetenz im Deutschunterricht am Beispiel der Hierarchie-Ebenen „lokale Kohärenzbildung“ und „Bildung von Superstrukturen“

Förderung der Lesekompetenz im Deutschunterricht

Richter und Christmann (2002) unterscheiden in ihrer Ausgestaltung eines kognitionstheoretischen Modells des Lesens verschiedene Hierarchie- oder Anforderungs-Ebenen (vgl. Tabelle 9), welche jeweils zum Ausgangspunkt für die Förderung der Lesekompetenz werden können (IQB, 2009; vgl. auch Rosebrock & Nix, 2008). Exemplarisch seien hier die Ebenen „lokale Kohärenzbildung“ als eine hierarchieniedrige und „Bildung von Superstrukturen“ als eine hierarchiehohe Prozessebene herausgegriffen. Unter „lokaler Kohärenzbildung“ werden Prozesse verstanden, bei denen die Schülerinnen und Schülern aus einzelnen Satzfolgen Sinnzusammenhänge bilden. Dies geschieht unter Berücksichtigung semantischer und syntaktischer Gesichtspunkte des Textes. Kontextspezifisches Vorwissen erleichtert es dabei Inferenzen zu bilden, also Leerstellen im Text sinnvoll zu ergänzen. Auf dieser hierarchieniedrigen Ebene können Viellese-Verfahren ein Baustein zur Förderung der Leseroutine und Lesemotivation sein (vgl. Rodebrock & Nix, 2008): Im Unterricht werden verbindliche Termine festgelegt, zu denen die Schülerinnen und Schüler individuell ausgewählte und im Unterricht nicht behandelte Texte der Kinder- und Jugendliteratur lesen. Schwache Leserinnen und Leser, die über keine Lesemotivation verfügen, werden durch diese Verpflichtung an das Lesen herangeführt. Geübtere Leserinnen und Leser können ohne Leistungsdruck eigenen Lesewünschen nachgehen, woraus zusätzliche Lesemotivation erwächst.

schwierigkeitsbestimmendes Merkmal		Ausprägungsgrad			
		sehr gering	gering	eher hoch	hoch
Komplexität der Aufgabenstellung	• Integrationsgrad (Grad der Menge und Dichte notwendiger Schlussfolgerungen	☐	☐	☐	☐
	• Implizitheitsgrad der Operatoren / der Arbeitsanweisungen	☐	☐	☐	☐
Komplexität und Anforderungshöhe des Bezugsgegenstandes (Text/Textensemble/ Problemstellung)	• Diversität des Bezugsgegenstands	☐	☐	☐	☐
	• Anspruchsniveau des Bezugsgegenstands	☐	☐	☐	☐
Anforderungen an Schülervoraussetzungen bzgl.	• Weltwissen	☐	☐	☐	☐
	• Fachwissen	☐	☐	☐	☐
Anforderungen an die sprachliche Darstellung der Produkte bzgl.	• Sprachliche Gestaltung (Lexik, Syntax)	☐	☐	☐	☐
	• Vielfalt der erwarteten Produktmerkmale	☐	☐	☐	☐
Umfang und Komplexität der notwendigen Reflexion und Bewertung	• Differenziertheit des Urteils	☐	☐	☐	☐
	• Verknüpfungsgrad der Bewertungsaspekte	☐	☐	☐	☐
	• Grad des Rückbezugs auf Bewertungsmaßstäbe	☐	☐	☐	☐

Abbildung 13: Analyseschema für schwierigkeitsbestimmende Merkmale von Lernaufgaben des Kompetenzbereichs „Leseverstehen“ im Fach Deutsch

Um auch eine Verbesserung der Lesekompetenz zu erzielen, ist das Viellesen allein möglicherweise nicht ausreichend (vgl. Rosebrock & Nix, 2008). Angesichts der Bedeutung des Vorwissens sollte daher neben dem textsortenbezogenen Wissen auch grammatisches und lexikalisches Wissen im Unterricht gezielt angesprochen werden. Eine Möglichkeit dazu bietet etwa die Wortschatzarbeit in einer Unterrichtssequenz mit idiomatischen Wendungen (vgl. IQB, 2009):

1. Zu Beginn der Unterrichtssequenz wird beispielhaft ein ausgewähltes Sprichwort behandelt.
2. Davon ausgehend diskutiert die Klasse allgemein die Bedeutung von Sprichwörtern.
3. Im Sinne der Wortschatzerweiterung analysieren die Schülerinnen und Schüler eine bestimmte Anzahl von Sprichwörtern in Gruppenarbeit. Die Gruppen bereiten eine visuelle Darstellung ihrer Ergebnisse vor.
4. Schließlich präsentieren die Gruppen ihre Arbeit der Klasse und beziehen diese in die Analyse ein.

Ebene	didaktisch beeinflussbare Einflussfaktoren	ausgewählte Fördermöglichkeiten
Buchstaben-, Wort- und Satzerkennung	Lesegenauigkeit (De- und Rekodierung von Wörtern)	intensive Dekodierübungen auf der Wortebene
	Automatisierung der Dekodierfähigkeit	Übungen zur Lautbewusstheit und Lautsysnthese
		Lautleseverfahren zur Verbesserung der Leseflüssigkeit und der Fähigkeit zum Sequenzieren von Sätzen
		Wortschatzarbeit
Lokale Kohärenzbildung auf der Satzebene	Bereichsspezifisches Vorwissen	Viellese-Verfahren (freie Lesezeiten)
	Semantische und syntaktische Analysefähigkeit der Sätze	Wortschatzarbeit
Globale Kohärenzbildung	Strategiewissen	Trainieren von Lesestrategien
	Selbsteinschätzung als Leser	
	Motivation	
Bildung von Superstrukturen	Sprach-, Text- und Weltwissen	Literaturunterricht
	Domänenspezifisches Wissen	
Erkennen rhetorischer Strategien	Fähigkeit zur Reflektion der Textstrukturen und rhetorischen Mittel	Produktorientierte Verfahren
Lese- und Lernmotivation	Selbsteinschätzung als Leser	Arbeit mit Selbsteinschätzungsverfahren
	Übernahme von Verantwortung für das eigene Lernen	Lese-Lerntagebuch
	Motivation	Verfahren der Leseanimation

Tabelle 9: Ebenen des Leseprozesses, didaktisch beeinflussbare Einflussfaktoren und ausgewählte Fördermöglichkeiten (vgl. IQB, 2009, S. 69)

Auf hierarchiehöherer Ebene steht u.A. die „Bildung von Superstrukturen“ im Fokus: Unter Superstrukturen sind Regeln und Kategorien zur Ordnung von Texten zu verstehen. Da sich Leserinnen und Leser bei einem neuen Text an ihrem bereits bestehenden Wissen (über die Eigenschaften von Textsorten) orientieren, besitzen Superstrukturen große Bedeutung für den Leseprozess. Die Fähigkeit zur Bildung von Superstrukturen ist wesentlich vom Sprach-, Text-, Welt- und domänenspezifischen Wissen abhängig. Sie kann im Unterricht insbesondere durch die Vermittlung von Wissen über Textsorten gefördert werden. Schülerinnen und Schüler müssen dazu die Möglichkeit erhalten, im Unterricht charakteristische Textstrukturen zu erarbeiten, wobei der Literaturunterricht bereits wesentliche Arbeit leistet. Unterentwickelt scheint bisher allerdings die Vermittlung von Textstrukturcharakteristika bei Sachtexten zu sein, welche sich deutlich von jenen bei narrativen Texten unterscheiden können. Hier ist die Weiterarbeit mit den Testmaterialien und -aufgaben zu empfehlen (vgl. IQB, 2009), indem einem Sachtext aus den Aufgaben der Vergleichsarbeiten ein weiterer Sachtext zum selben oder zu einem ähnlichen Thema zur Seite gestellt wird. Im Unterricht wird dann textvergleichend vorgegangen, indem in Gruppenarbeit der Aufbau beider Texte schematisch gegenübergestellt wird. Die Schülerinnen und Schüler arbeiten so charakteristische Muster heraus.

Beispiel 2: Von der Testaufgabe zur Lernaufgabe – Förderung der Modellierungskompetenz im Mathematikunterricht

Förderung der Modellierungskompetenz im Mathematikunterricht

Die nachfolgend beschriebene Aufgabe aus der Leitidee „Daten und Zufall“ wurde als Modifikation einer Aufgabe der Bildungsstandards formuliert (Lankes et al., 2005). Sie ermöglicht es darzustellen, wie unter demselben Stimulus durch Öffnung der Aufgabenformate verschiedene Einsatzmöglichkeiten als Test- und als Lernaufgabe im Unterricht konstruiert werden können. Unter den in Kapitel 2.4.2.3 aufgeführten Kompetenzen steht *Modellieren* im Zentrum; die Anforderungsbereiche hängen vom jeweiligen Aufgabenteil ab:

Weitspringen, der Springer-Cup

„An der Erich-Hüpf-Schule wird jedes Jahr eine 6. Klasse mit dem „Springer-Cup“ ausgezeichnet. Den Preis erhält die Klasse mit den besten Weitsprungergebnissen. Drei Klassen gehen dieses Mal mit folgenden Weiten ins Rennen (Abbildung 14).“

Sprungweiten (auf 10 cm gerundet)

Klasse 6a	Klasse 6b	Klasse 6c
390	390	370
380	370	380
380	390	400
390	400	390
400	390	410
390	360	410
390	370	360
380	430	380
390	400	370
380	380	360
380	370	400
390	390	390
390	380	410
370	410	390
390	390	390
380	380	370
390	390	380
390	370	410
380	380	360
390	370	370
380	390	380
380		400
390		390

In der Jury gibt es verschiedene Vorschläge, wie der Preis vergeben werden soll:

Jury A
„Die Sache ist doch klar, den Preis bekommt die Klasse mit dem Spitzenspringer."

Jury B
„Mit dem Preis soll doch die gesamte Klasse ausgezeichnet werden und nicht nur der beste Springer. Also gehört der Preis der Klasse, die die meisten guten Springer besitzt."

Jury C
„Und was, wenn der Rest grottenschlecht ist? Der ausgeglichensten Klasse gehört der Preis."

Abbildung 14: Stimulus der Aufgabe „Weitspringen, der Springer-Cup" in Anlehnung an Lankes et al. (2005, S. 37)

Aufbauend auf diesem Stimulus lassen sich einerseits Testaufgaben zur Kompetenzbestimmung formulieren, wie sie sich etwa bei Vergleichsarbeiten finden, z.B.

1. Bestimme das arithmetische Mittel und den Zentralwert für jede Klasse. Was lässt sich feststellen?
2. Welche Klasse bekäme den Preis, wenn die Vergabe nach der Regelung von A, B oder C vorgenommen werden würde? Begründe jeweils.
3. An welche Klasse sollte deiner Meinung nach der „Springer-Cup" vergeben werden? Schreibe einen Bericht an die Jury, in dem du deinen Vergabemodus begründest.

Darüber hinaus kann die Aufgabenstellung zu einer Lernaufgabe umformuliert werden, etwa:

An welche Klasse sollte eurer Meinung nach der Springer-Cup vergeben werden?

Schreibt einen Bericht, in dem ihr eure Meinung deutlich macht. Folgende Bereiche sollten darin angesprochen werden: Arithmetischer Mittelwert; Zentralwert; Anfertigen von Säulendiagrammen; Umgang mit Diagrammen und Kennwerten.

Diese Formulierung bietet Schülerinnen und Schüler die folgenden Möglichkeiten zum Kompetenztraining:

1. Sie bearbeiten eine alltagsnahe mathematische Fragestellung eigenverantwortlich entsprechend ihrer mathematischen Kompetenzen. Dabei spielen die rechnerischen Anforderungen gegenüber der verständnisvollen Entwicklung mathematischer Konzepte eine eher untergeordnete Rolle, auch wenn die Aufgabe durchaus umfangreiche Rechnungen verlangt (Einsatz des Taschenrechners möglich).
2. Eine arbeitsteilige Bewältigung im Rahmen einer Gruppenarbeit ist möglich. Da sich bei den ausgewählten Ergebnissen sowohl durch den arithmetischen Mittelwert als auch den Zentralwert keine klaren Sieger festlegen lassen, sind bei einer Gruppenarbeit unterschiedliche Vorschläge zu erwarten. Die Schülerinnen und Schüler erleben, dass die Entscheidung dann nur normativ in der Diskussion getroffen werden kann.
3. Es besteht die Möglichkeit, eine offene Aufgabenstellung zu formulieren, bei der die Datenaufbereitung von den Schülerinnen und Schüler selbst übernommen wird, die dabei verschiedene Phasen des Modellierungskreislaufs durchlaufen.
4. Eine Modifikation der Aufgabe, beispielsweise auf selbstständig erhobene Daten, ist genauso denkbar wie fächerübergreifendes Arbeiten in Kooperation mit anderen Fächern.

An den beiden vorgestellten Beispielen wird deutlich, dass die Durchführung, Auswertung und Interpretation von Vergleichsarbeiten alleine noch nicht zur Verbesserung des Unterrichts ausreichen, die eigentliche Arbeit vielmehr noch folgt. Sie verdeutlichen allerdings auch: Vergleichsarbeiten können ein Ausgangspunkt für die Entwicklung gehaltvoller Konzepte zum kompetenzorientierten Unterricht sein.

2.5.5 Weiterführende Literatur

Literatur

Die Nutzung von Ergebnissen aus Vergleichsarbeiten für die Unterrichtsentwicklung beschreiben Peek & Dobbelstein (2006) sowie Bonsen, Büchter & Peek (2006). Zu empfehlen ist das Herausgeberwerk von Kuper & Schneewind (2006) zur Rezeption von Schulleistungsergebnissen, in dem sich unter anderem Groß-Ophoff, Koch, Hosenfeld & Helmke mit der Rezeption der Ergebnisrückmeldungen aus dem VERA-Projekt befassen. Eine systematische Aufarbeitung empirischer Erkenntnisse zur Ergebnisrezeption und -nutzung hat zuletzt Dedering (2011) vorgestellt; ihre Analyse weist auf Optimierungsbedarf bei schulischen Aktivitäten im Anschluss an die Ergebnisrückmeldung hin. Anregungen zur Verwendung der Ergebnisrückmeldung für Prozesse der Unterrichtsentwicklung geben die Webseiten der Ministerien. Die hier vorgestellten Ideen zur Förderung der Lesekompetenz knüpfen an Rosebrock & Nix (2008) an. Der Förderung von Lesekompetenz hat sich das Projekt ProLesen (http://www.leseforum.bayern.de/index.asp?MNav=6) verschrieben. Vorschläge zur Entwicklung kompetenzorientierter Mathematikaufgaben bieten z.B. Blum, Drüke-Noe, Hartung & Köller (2006) sowie Büchter & Leuders (2005).

2.5.6 Verständnis und Diskussionspunkte

1. *Beschreiben Sie bitte, wie die unter 2.5.1 aufgeführten Informationen der Ergebnisrückmeldung genutzt werden können. Für welchen Zweck wird die jeweilige Information zur Verfügung gestellt?*
2. *Sie erhalten im Rahmen von Vergleichsarbeiten die folgende Rückmeldung über das Abschneiden der Jahrgangsstufe:*

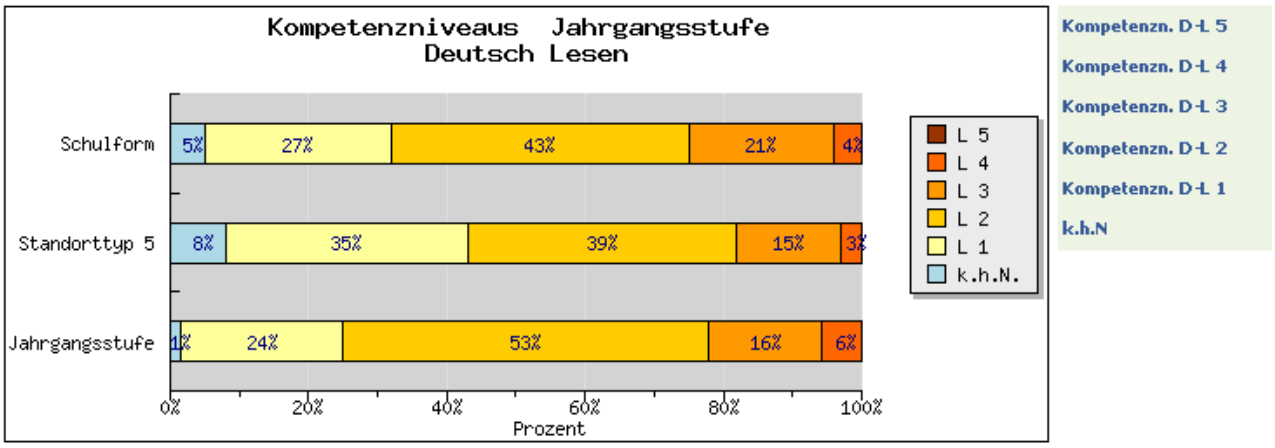


Abbildung 15: Beispiel einer Ergebnisrückmeldung: Verteilung einer Jahrgangsstufe und Referenzverteilungen auf fünf Kompetenzstufen zuzüglich des nicht auswertbaren Bereiches (k.h.N. – „Kein hinreichender Nachweis für das Erreichen eines Kompetenzniveaus“)

Welche Aussagen sind Abbildung 15 zu entnehmen?

3. *Im Mittelpunkt der Darstellung dieses letzten Kapitels stand die Zielsetzung, ausgehend von der Ergebnisrückmeldung aus Vergleichsarbeiten Unterrichtsentwicklung zu betreiben. Skizzieren Sie bitte kurz einige Möglichkeiten, Ergebnisrückmeldungen zur Verbesserung der diagnostischen Kompetenzen von Lehrkräften zu nutzen.*

2.6 Literatur

Arnold, Karl-Heinz (1999*); Fairneß bei Schulsystemvergleichen; Diagnostische Konsequenzen von Schulleistungsstudien für die unterrichtliche Leistungsbewertung und binnenschulische Evaluation*; Münster, Waxmann

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001); Lesekompetenz: Testkonzeption und Ergebnisse; In J. Baumert, E. Klieme, M. Neubrand, Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J. & Weiß, M. (Hrsg.), *PISA 2000: Basiskompetenzen von Schülern im internationalen Vergleich* (S. 69-137); Opladen, Leske & Budrich

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

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Tillmann, K.-J. & Prenzel, M. (Hrsg.) (2003); *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland*; Opladen, Leske + Buderich

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

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

Baumert, J., Trautwein, U. & Artelt, C. (2003); Schulumwelten – institutionelle Bedingungen des Lehrens und Lernens; In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann, & M. Prenzel (Hrsg.); *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 259-330); Opladen, Leske + Buderich

Blum, W., Drüke-Noe, C., Hartung, R. & Köller, O. (Hrsg.) (2006); *Bildungsstandards Mathematik: konkret;* Berlin, Cornelsen Skriptor

Bond, T.G. & Fox, C.M. (2001); *Applying the Rasch Model: fundamental measurement in the human sciences;* Mahwah, NJ, LEA

Bonsen, M., Büchter, A. & Peek, R. (2006); Datengestützte Schul- und Unterrichtsentwicklung – Bewertungen der Lernstandserhebungen in NRW durch Lehrerinnen und Lehrer; In W. Bos, H.G. Holtappels, H. Pfeiffer, H.-G. Rolff, & R. Schulz-Zander (Hrsg.). *Jahrbuch der Schulentwicklung* (Vol. 14; S. 125-148); Weinheim, Juventa

Bos, W., Hornberg, S., Arnold, K.-H., Faust,G., Fried, L., Lankes, E.-V., Schwippert, K. & Valtin, R. (Hrsg.) (2007); *IGLU 2006 – Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich;* Münster, Waxmann

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

Bremerich-Vos, A., Granzer, D. & Köller, O. (Hrsg.) (2008); Lernstandsbestimmung im Fach Deutsch. Gute Aufgaben für den Unterricht*;* Weinheim, Beltz

Bremerich-Vos, A., Granzer, D., Behrens, U. & Köller, O. (Hrsg.) (2009); *Bildungsstandards für die Grundschule: Deutsch konkret: Aufgabenbeispiele – Unterrichtsanregungen – Fortbildungsideen;* Berlin, Cornelsen Skriptor

Büchter, A. & Leuders, T. (2005). *Mathematikaufgaben selbst entwickeln. Lernen fördern – Leistung prüfen*. Berlin, Cornelsen Scriptor

Burkard, C. & Peek, R. (2004); Anforderungen an zentrale Lernstandserhebungen; Ein Werkstattbericht aus Nordrhein-Westfalen; *Pädagogik*, 6, 24-27

Dedering, K. (2011). Hat Feedback eine positive Wirkung? Zur Verarbeitung extern erhobener Leistungsdaten in Schulen; *Unterrichtswissenschaft*, 39, 63-83

Dobbelstein, P. & Peek, R. (2007); Einleitung: Lernstandserhebungen als Beitrag zu einer empiriegestützten Unterrichtsentwicklung; In MSW (Hrsg.), *Lernstandserhebungen Mathematik in Nordrhein-Westfalen; Impulse zum Umgang mit zentralen Tests* (S. 7-13); Stuttgart, Klett

DESI-Konsortium (Hrsg.) (2008); *Unterricht und Kompetenzerwerb in Deutsch und Englisch; Ergebnisse der DESI-Studie*; Weinheim, Beltz

Eikenbusch, G. & Leuders, T. (Hrsg.) (2004); *Lehrer-Kursbuch Statistik*; Berlin, Cornelsen Scriptor

Europarat (2001); *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*; Berlin, Langenscheidt

Fleischer, J., Wirth, J. & Leutner, D. (2007); Testmethodische Grundlagen der Lernstandserhebungen NRW: Erfassung von Schülerkompetenzen für Vergleiche mit kriterialen und sozialen Bezugsnormen; In MSW (Hrsg.), *Lernstandserhebungen Mathematik in Nordrhein-Westfalen; Impulse zum Umgang mit zentralen Tests* (S. 91-113); Stuttgart, Klett

Freudenthal, H. (1977); *Mathematik als pädagogische Aufgabe (Bd. 1, Bd. 2);* Stuttgart, Ernst Klett Verlag

Freudenthal, H. (1983); *Didactical phenomenology of mathematical structures*; Dordrecht, Reidel

Groß Ophoff, J., Koch, U., Hosenfeld, I. & Helmke, A. (2006); Ergebnisrückmeldungen und ihre Rezeption im Projekt VERA; In H. Kuper & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen – Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 19-40); Münster, Waxmann

Helmke, A. & Hosenfeld, I. (2003a); Vergleichsarbeiten (VERA): eine Standortbestimmung zur Sicherung schulischer Kompetenzen – Teil 1: Grundlagen, Ziele, Realisierung; *SchulVerwaltung NRW*, 4, 107-110

Helmke, A. & Hosenfeld, I. (2003b); Vergleichsarbeiten (VERA): eine Standortbestimmung zur Sicherung schulischer Kompetenzen – Teil 2: Nutzung für Qualitätssicherung und Verbesserung der Unterrichtsqualität; *SchulVerwaltung NRW*, 5, 143-145

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

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2004); Vergleichsarbeiten als Werkzeug für die Verbesserung der diagnostischen Kompetenz von Lehrkräften; In R. Arnold & C. Griese (Hrsg.); *Schulleitung und Schulentwicklung* (S. 119-144); Hohengehren, Schneider

Helmke, A. & Schrader, F. W. (2001); Determinanten der Schulleistung; In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (2. Aufl., S. 81-91); Weinheim, Beltz

IQB (ohne Jahr); *Deutsch Sekundarstufe I*; Verfügbar unter: https://www.iqb.hu-berlin.de/arbbereiche/projekte/?pg=p_36 [24.07.2009]

IQB (2009); *Deutsch: VERA 8 Handreichung 2009 – Testheft II.* Verfügbar unter: http://www.iqb.hu-berlin.de/vera2?reg=r_6 [24.07.2009]

IQB (2010); *Vergleichsarbeiten 2010. 3. Jahrgangsstufe (VERA-3). Mathematik – Didaktische Handreichung*. Verfügbar unter: http://www.standardsicherung.schulministerium.nrw.de/vera3/upload/download/mat_10-11/VERA_M_Did_Hand_Mathematik.pdf

Isaac, K. (2011); Neues Standorttypenkonzept. Faire Vergleiche bei Lernstandserhebungen. *Schule NRW 06/11. Amtsblatt des Ministeriums für Schule und Weiterbildung*, 300-301

Isaac, K., Metzeld, D. & Eichler, W. (2009); Bewusster Umgang mit Sprache – Sprache und Sprachgebrauch untersuchen; *Grundschulunterricht*, *56 (2),* 28-31

Klauer, K.J. (1987); *Kriteriumsorientierte Tests*; Göttingen, Hogrefe

Kliemann S. (Hrsg.) (2010); *Diagnostizieren und Fördern – Kompetenzen erkennen, unterstützen und erweitern. Beispiele und Anregungen für die Jahrgänge 1 bis 4;* Berlin, Cornelsen Verlag Scriptor

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

Klieme, E., Hartig, J. & Rauch, D. (2008); The concept of competence in educational contexts; In J. Hartig, E. Klieme, D. Leutner (Hrsg); *Assessment of competencies in educational contexts* (S. 3-22); Göttingen, Hogrefe & Huber

Klieme, E. & Leutner, D. (2006); Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen; Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG; *Zeitschrift für Pädagogik*, 52, 876-903

KMK (2003a); *Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss: Beschluss vom 04. 12. 2003*; München, Luchterhand

KMK (2003b); *Bildungsstandards für die erste Fremdsprache (Englisch / Französisch) für den Mittleren Schulabschluss: Beschluss vom 04. 12. 2003*; München, Luchterhand

KMK (2003c); *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss: Beschluss vom 04. 12. 2003*; München, Luchterhand

KMK (2004a); *Bildungsstandards im Fach Deutsch für den Primarbereich: Beschluss vom 15. 10. 2004*; München, Luchterhand

KMK (2004b); *Bildungsstandards im Fach Mathematik für den Primarbereich: Beschluss vom 15. 10. 2004*; München, Luchterhand

KMK (2004c); *Bildungsstandards im Fach Deutsch für den Hauptschulabschluss: Beschluss vom 15. 10. 2004*; München, Luchterhand

KMK (2004d); *Bildungsstandards für die erste Fremdsprache (Englisch / Französisch) für den Hauptschulabschluss: Beschluss vom 15. 10. 2004*; München, Luchterhand

KMK (2004e); *Bildungsstandards im Fach Mathematik für den Hauptschulabschluss: Beschluss vom 15. 10. 2004*; München, Luchterhand

KMK & for.mat (ohne Jahr); *Fortbildungskonzepte und -materialien zur kompetenz- bzw. standardbasierten Unterrichtsentwicklung*. Verfügbar unter: http://www.kmk-format.de/ [22.07.2010]

KMK & IQB (2006); *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*; München, Luchterhand

Köller, O. (2008); Bildungsstandards in Deutschland: Implikationen für die Qualitätssicherung und Unterrichtsqualität; *Zeitschrift für Erziehungswissenschaft*, *Sonderheft 9*, 47-59

Köller, O. & Baumert, J. (2002); Entwicklung von Schulleistungen; In R. Oerter & L. Montada (Hrsg.); *Entwicklungspsychologie* (5. Auflg., S. 756-786); Weinheim, Beltz/PVU

Kuper, H. & Schneewind, J. (2006); *Rückmeldung und Rezeption von Forschungsergebnissen; Zur Verwendung wissenschaftlichen Wissens im Bildungsbereich*; Münster, Waxmann

Lankes, E.-M., Lorenzen, H., Petersen, C., von Urban, S. & Zielinski, D. (2005). *Kompetenzorientierter Mathematikunterricht. Anregungen für die Arbeit mit den Bildungsstandards zum Hauptschulabschluss und mittleren Abschluss (Sekundarstufe I).* Kronshagen: Institut für Qualitätsentwicklung an Schulen.

Lehmann, R. H. & Peek, R. (1997); *Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Klassen an Hamburger Schulen; Bericht über die Untersuchung im September 1996 (unveröffentlichter Forschungsbericht)*; Hamburg

Leutner, D. (2010); Pädagogisch-psychologische Diagnostik; In D. H. Rost (Hrsg.); *Handwörterbuch Pädagogische Psychologie* (4. überarbeitete Aufl., S. 624-635); Weinheim, PVU

Leutner, D., Fleischer, J., Spoden, C. & Wirth, J. (2007); Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik; *Zeitschrift für Erziehungswissenschaft, Sonderheft 8*, 149-167

Lienert, G.A. & von Eye, A. (1998); *Erziehungswissenschaftliche Statistik; Eine elementare Einführung für pädagogische Berufe*; Weinheim, Beltz

Möller, G., Pallack, A. & Fleischer, J. (2007); Da schau hin: Was Lehrerinnen und Lehrer aus Lernstandserhebungen über ihre schwachen Schülerinnen und Schüler erfahren können; In A. Peter-Koop & A. Bikner-Ahsbahs (Hrsg.); *Mathematische Bildung – Mathematische Leistung; Festschrift für Michael Neubrand zum 60. Geburtstag* (S. 97-113). Hildesheim, Franzbecker

Möller, J. & Schiefele, U. (2004); Motivationale Grundlagen der Lesekompetenz; In U. Schiefele, C. Artelt, W. Schneider, & P. Stanat (Hrsg.); *Entwicklung, Bedingungen und Förderung der Lesekompetenz: Vertiefende Analysen der PISA-2000-Daten* (S. 101-124); Wiesbaden, Verlag für Sozialwissenschaften

Nachtigall, C. & Jantowski, A. (2004); Die Thüringer Kompetenztests; *Neue Praxis der Schulleitung, Thüringen*, 73, 1-14

Nachtigall, C., Kröhne, U., Enders, U. & Steyer, R. (2008); Considering the influence of context variables on students competencies; In J. Hartig, E. Klieme & D. Leutner (Hrsg.); *Assessment of Competencies in Educational Contexts* (S. 315-335); Göttingen, Hogrefe & Huber

Nachtigall, C., Kröhne, U. & Müller, M. (2005); *Ergänzungen zum Ergebnisbericht der Kompetenztests 2005;* Verfügbar unter: http://www.kompetenztest.de/download/kt05/Muster_Ergaenzg_MK06.pdf [24.07.2009]

Netzwerk Empiriegestützte Schulentwicklung (2006); *Positionspapier zu: Zentrale standardisierte Lernstandserhebungen;* Verfügbar unter: http://www.iqb.hu-berlin.de/bista/dateien/EMSE_Positionsp.pdf [24.07.2009]

Peek, R. (2004); Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik (Qua-SUM)“ – Klassenbezogene Ergebnisrückmeldungen und ihre Rezeption in Brandenburger Schulen; *Empirische Pädagogik*, 18 (Themenheft), 82-114

Peek, R. & Dobbelstein, P. (2006); Zielsetzung: Ergebnisorientierte Schul- und Unterrichtsentwicklung; Potenziale und Grenzen der nordrhein-westfälischen Lernstandserhebungen; In Böttcher, W., Holtappels, H. G. & Brohm, M. (Hrsg.); *Evaluation im Bildungswesen; Eine Einführung in Grundlagen und Praxisbeispiele* (Grundlagentexte Pädagogik, S. 177-194); Weinheim und München, Juventa

Projekt VERA (2008). *Didaktische Erläuterungen „Leseverständnis“ und „Sprache und Sprachgebrauch untersuchen“.* Verfügbar unter: http://139.14.28.6/verapub/fileadmin/downloads/2008/VERA_D_didakt_Erlaeut_2008.pdf [28.01.2010]

Projekt VERA (2009); *Korrekturanweisungen für die Deutschaufgaben 2009.* Verfügbar unter: http://139.14.28.6/verapub/fileadmin/downloads/2009/VERA_D_Korrekturanweisung_2009.pdf [28.01.2010]

Ramm, G., Prenzel, M., Heidemeier, H. & Walter, O. (2004); Soziokulturelle Herkunft: Migration; In Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J. & Schiefele, U.(Hrsg.); *Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (S. 254-272); Münster, Waxmann

Rasch, G. (1960); *Probabilistic models for some intelligence or attainment tests;* Copenhagen, Nielsen & Lydiche (2nd Edition Chicago University of Chicago Press, 1980)

Reiss, K. & Winkelmann, H. (2009); Kompetenzstufenmodelle für das Fach Mathematik im Primarbereich; In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 120-141); Weinheim, Beltz

Richter, T. & Christmann, U. (2002); Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz: Bedingungen, Dimensionen, Funktionen* (S. 25-59); Weinheim, Juventa

Rosebrock, C. & Nix, D. (2008); *Grundlagen der Lesedidaktik und der systematischen Leseförderung*. Baltmannsweiler, Hohengehren

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

Ruep, M & Keller, G. (2007); *Schulevaluation*; Frankfurt, Peter Lang

Rupp, A. A., Vock, M., Harsch, C. & Köller, O. (2008); *Developing standards-based assessment items for English as a first foreign language – Context, processes, and outcomes in Germany;* Münster, Waxmann

Schnotz, W. & Dutke, S. (2004); Kognitionspsychologische Grundlagen der Lesekompetenz: Mehrebenenverarbeitung anhand multipler Informationsquellen; In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.); *Struktur, Entwicklung und Förderung von Lesekompetenz; Vertiefende Analysen im Rahmen von PISA 2000* (S. 61-99); Wiesbaden, VS Verlag für Sozialwissenschaften

Schrader, F.-W. (2001); Diagnostische Kompetenz von Eltern und Lehrern; In D.H. Rost (Hrsg.); *Handwörterbuch Pädagogische Psychologie* (2. überarb. u. erw. Aufl., S. 91-96); Weinheim, Beltz

Schrader, F.-W. & Helmke, A. (2001); Alltägliche Leistungsbeurteilung durch Lehrer; In F. E. Weinert (Hrsg.); *Leistungsmessungen in Schulen* (S. 45-58); Weinheim, Beltz

Schräpler, J.-P. (2009); Verwendung von SGB II-Dichten als Raumindikator für die Sozialberichterstattung am Beispiel der „sozialen Belastung" von Schulstandorten in NRW – ein Kernel-Density-Ansatz. *Statistische Analysen und Studien Nordrhein-Westfalen, 57*, 3-28

Spinath, B. (2005); Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer/innen und das Konstrukt der diagnostischen Kompetenz; *Zeitschrift für Pädagogische Psychologie, 19,* 85-95

Süß, H. M. (2001); Prädiktive Validität der Intelligenz im schulischen und außerschulischen Bereich; In E. Stern & J. Guthke (Hrsg.); *Perspektiven der Intelligenzforschung* (S. 109-135); Lengerich, Pabst Science Publishers

Tenorth, H.-E. (2001); Englisch: Ein Kerncurriculum, seine Notwendigkeit und seine Gestalt – Zusammenfassung; In H.-E. Tenorth, (Hrsg.); *Kerncurriculum Oberstufe; Mathematik – Deutsch – Englisch; Expertisen im Auftrag der Ständigen Konferenz der Kultusminister* (S. 156-161); Weinheim, Beltz

Tesch, B., Leupold, E., Köller, O. (Hrsg.) (2008); *Bildungsstandards Französisch: konkret*; Berlin, Cornelsen Scriptor

Walther, G., van den Heuvel-Panhuizen, M., Granzer, D. & Köller, O. (2007); *Bildungsstandards für die Grundschule: Mathematik konkret*; Berlin, Cornelsen Skriptor

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

Wirtz, M. & Caspar, F. (2002); *Beurteilerübereinstimmung und Beurteilerreliabilität*; Göttingen, Hogrefe

Ziener, G. (2006); *Bildungsstandards in der Praxis. Kompetenzorientiert unterrichten*; Seelze, Velber

Bildungsmonitoring auf der Systemebene

Nina Hovenga
Wilfried Bos

UDiKom

Aus- und Fortbildung der Lehrkräfte in Hinblick auf Verbesserung der Diagnosefähigkeit, Umgang mit Heterogenität, individuelle Förderung

Bildungsmonitoring auf der Systemebene

Alle im Projekt erstellten Materialien
finden Sie unter
www.udikom.de

3. Einleitung

Der vorliegende Studienbrief-Teil beschäftigt sich mit dem Systemmonitoring als Teil des Bildungsmonitorings[1]. Dabei wird unter Bildungsmonitoring in der Literatur verstanden:

Definition: Bildungsmonitoring

„Bildungsmonitoring meint dabei die kontinuierliche und systematische, auf wissenschaftliche Methoden gestützte Beobachtung der Bedingungen, Verläufe, Ergebnisse und Wirkungen von Bildungsprozessen in und außerhalb von Institutionen mit dem Ziel, bildungspolitischen Akteuren wissenschaftlich aufbereitetes, empirisches Wissen zur Information für bildungspolitische Entscheidungsprozesse zur Verfügung zu stellen." (Wolter, 2009)

„Bildungsmonitoring ist die systematische und auf Dauer angelegte Beschaffung und Aufbereitung von Informationen über ein Bildungssystem und dessen Umfeld." (Maritzen, 2008)

Im Jahr 2006 wurde durch die KMK eine Gesamtstrategie zum Bildungsmonitoring beschlossen. Bildungsprozesse in Deutschland sollen durch vier miteinander verbundene Bereiche beobachtet und weiterentwickelt werden:

- regelmäßige Teilnahme an internationalen Schulleistungsuntersuchungen
- zentrale Überprüfung der Bildungsstandards im Ländervergleich
- Vergleichsarbeiten
- gemeinsame Bildungsberichterstattung

Das übergeordnete Ziel dieser Gesamtstrategie liegt in der Beschaffung von Informationen, die für die Steuerung des Bildungssystems und für die Schul- und Unterrichtsentwicklung benötigt werden. Zudem sollen die gewonnenen Erkenntnisse mit Maßnahmen der Qualitätsentwicklung verknüpft werden, sodass die pädagogische Arbeit an jeder einzelnen Schule profitieren kann (KMK, 2006).

Wichtig für die Durchführung des Bildungsmonitorings ist es, dass Wissenschaft, Bildungspolitik und Bildungspraxis miteinander arbeiten (Stadelmann, 2008). Durch die kombinierte Arbeit von Theorie und Praxis können Studien im Rahmen des Bildungsmonitorings nicht nur ‚bildungspolitischen Akteuren' sondern auch Lehrkräften bei ihrer Arbeit behilflich sein. Zudem gibt es viele Berührungspunkte einer Lehrkraft mit dem Bildungsmonitoring. Aus diesen Punkten folgt, dass es auch für Lehrkräfte von Bedeutung ist, über die Studien des Systemmonitorings Bescheid zu wissen, da sie einen Einfluss auf den Lehralltag haben (werden).

Hierzu zählen:

- Diskussionen in den Medien über die Wirksamkeit von Schule
- Diskussionen innerhalb des Kollegiums
- Eltern, die durch die erhöhte Transparenz der Schulwirksamkeit wissen wollen, ob ihre eigenen Kinder besser oder ähnlich abgeschnitten hätten
- Konsequenzen der Bildungsadministration:
  - Bildungsstandards (vgl. Kapitel 3.5.3.1)
  - Kernlehrpläne (vgl. Kapitel 3.5.3.2)
  - Sprachförderung (vgl. Kapitel 3.5.3.3)
  - Ganztagsschulen (vgl. Kapitel 3.5.3.4)
- Projekte (KMK, Ministerien, Schulamt, Stadt) (vgl. Kapitel 3.5.3.5 bis 3.5.3.7)
- eigene Schülerinnen und Schüler, Schülerinnen und Schüler der eigenen Schule oder Schülerinnen und Schüler von bekannten Lehrkräften werden getestet

Bildungsmonitoring / Evaluation

Von großer Bedeutung ist es, den Begriff des ‚Bildungsmonitorings' von dem der ‚Evaluation' abzugrenzen. Zwar werden auch im Rahmen des Bildungsmonitorings evaluative Aussagen über Teilpopulationen, beispielsweise im Vergleich der Bundesländer oder der Schulformen getroffen, dennoch soll der

1 Im Folgenden wird die Bezeichnung ‚Systemmonitoring als Teil des Bildungsmonitorings' verkürzt als ‚Systemmonitoring' verwendet.

Begriff ‚Evaluation' dafür reserviert werden, dass der „Erfolg einer bestimmten Maßnahme bzw. die Leistungsfähigkeit einer einzelnen Institution (z.B. einer Schule) zu beurteilen ist" (Klieme u.a., 2007). Die weiteren Unterschiede fasst Tabelle 1 zusammen.

	Bildungsmonitoring	Evaluation
Grundlage/Ziel	Informationen über das Bildungssystem	Erfolg bestimmter Maßnahmen/Institutionen
Untersuchungsgegenstand	Unterschiedliche Bereiche des Bildungssystems	Bestimmte Maßnahme
Untersuchungsart	Umfassend	Eng und detailliert
Ergebnis	Informationen	Konkrete Entscheidung

Tabelle 1: Abgrenzung der Begriffe ‚Bildungsmonitoring' und ‚Evaluation'

Durch die Unterschiede wird deutlich, dass eine Studie nicht gleichzeitig beiden ‚Anliegen' gerecht werden kann (Klieme u.a., 2007). Allerdings ist es möglich, im Rahmen des ‚normalen' Bildungsmonitorings ergänzende Befragungen mit evaluativem Charakter durchzuführen. Beispielsweise wurden in einigen Studien ergänzende Videoanalysen erstellt, um das Lehrverhalten zu betrachten, andere Studien betrachteten den Zusammenhang zwischen den Ergebnissen und dem sozialen Hintergrund der Schülerinnen und Schüler. Zudem wurden in den meisten Studien Zusatzstudien auf nationaler Ebene durchgeführt, sodass hierdurch untersucht werden konnte, ob beispielweise Bildungsstandards erreicht werden (Klieme u.a., 2007).

Der vorliegende Studienbrief-Teil befasst sich mit einem Teil des Bildungsmonitorings, dem Systemmonitoring.

Definition: Systemmonitoring

Systemmonitoring ist ein Teil des Bildungsmonitorings und bezeichnet „Maßnahmen der quantitativen Erfassung von Bildungserträgen auf Systemebene" (Stanat, 2008).

3.1 Zielsetzung

Kapitel 3.1 beschäftigt sich mit der Zielsetzung von Systemmonitoring. Dazu wird betrachtet, warum Studien des Systemmonitorings durchgeführt werden und welche Ziele sie im Einzelnen verfolgen. Zudem werden die vier Studien TIMSS[2], PISA, IGLU und DESI vorgestellt, die bekanntesten Studien im Rahmen des Systemmonitorings.

In diesem Kapitel werden folgende Fragen beantwortet:

- *Welche Charakteristika haben Systemmonitoring-Studien?*
- *Warum werden Systemmonitoring-Studien durchgeführt?*
- *Welche Studien gibt es und welche Inhalte haben diese?*

Bei den vier oben genannten Studien handelt es sich jeweils um Large-Scale Assessments.

Definition: Large-Scale Assessments

- Leistungsmessungen mit einer hohen Anzahl an Testpersonen
- Untersuchungen, die messen, welche Kompetenzen Personen einer bestimmten Zielgruppe haben
- internationaler Vergleich dieser Kompetenzen
- Stichprobenuntersuchungen, bei denen nur in einigen wenigen Schulen aus jedem Staat Schülerinnen und Schüler getestet werden
- Erhebungen in vielen verschiedenen Staaten
- Durchführung sehr genau geplant, in der Praxis nicht sehr flexibel

Datenrückmeldung an Schulen

Eine Rückmeldung der Ergebnisse an die an den Studien beteiligten Schulen ist nur eingeschränkt möglich. Vor allem die Vorgehensweise der Rückmeldung ist als nicht adressatengerecht kritisiert worden (Rolff, 2008), zudem erfolgt sie durch die komplexe und internationale Auswertung der Daten erst sehr spät. Auch ist die Bedeutung der Datenrückmeldung an die einzelnen Schulen, die an den Erhebungen teilgenommen haben, durch die Vergleichsarbeiten (vgl. Studienbrief-Teil Vergleichsarbeiten) zurückgegangen, da diese alle Klassen und alle Schülerinnen und Schüler betreffen.

Eine solche Rückmeldung an die Schulen ist aber auch nicht das bedeutsamste Ziel von Systemmonitoring-Studien.

3.1.1 Begründung für Systemmonitoring–Studien

Wichtigkeit Systemmonitoring

Systemmonitoring ist sinnvoll und notwendig, da nur so Informationen für die Steuerung des Bildungssystems gewonnen werden können. Nur auf diese Weise können sich Bildungsinstitutionen entwickeln und optimieren.

Abbildung 1 zeigt diesen Verlauf.

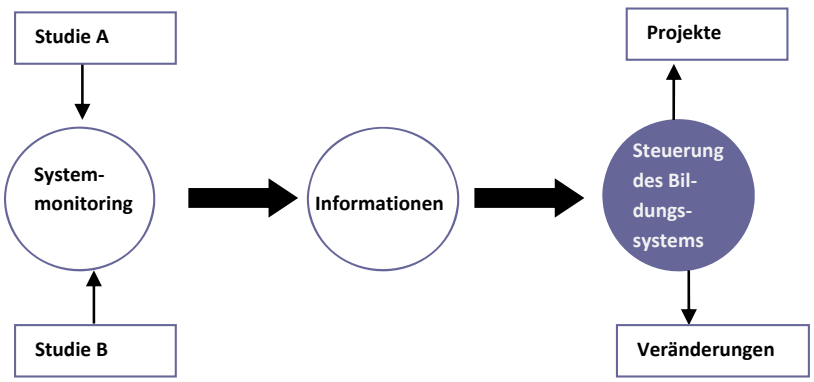


Abbildung 1: Funktionsweise des Systemmonitorings

2 Die Bedeutungen dieser Abkürzungen werden in den Kapiteln 3.1.3.1 bis 3.1.3.4 erklärt.

Die Abbildung 1 zeigt, dass sich das Systemmonitoring aus mehreren Projekten (hier: A und B) zusammensetzt. Aus diesen Studien ergeben sich dann Informationen, die für die Steuerung des Bildungssystems verwendet werden, beispielsweise um Projekte zu initiieren und Veränderungen zu bewirken.

Um zu verdeutlichen, warum das Systemmonitoring wichtig für den Schulalltag ist, zeigt Abbildung 2 den Einfluss auf den Unterricht, der zwischen der Lehrkraft und der individuellen Förderung der Schülerinnen und Schüler steht. Das Ziel des Unterrichts ist es, dass der Schüler individuell gefördert wird und einen Wissenszuwachs erreicht. Dazu erhält er einen Unterricht, der durch die Lehrkraft gestaltet wird. Aber nicht nur die Lehrkraft hat einen Einfluss auf den Unterricht, sondern auch die Projekte und Veränderungen, die sich aus der Steuerung des Bildungssystems ergeben.

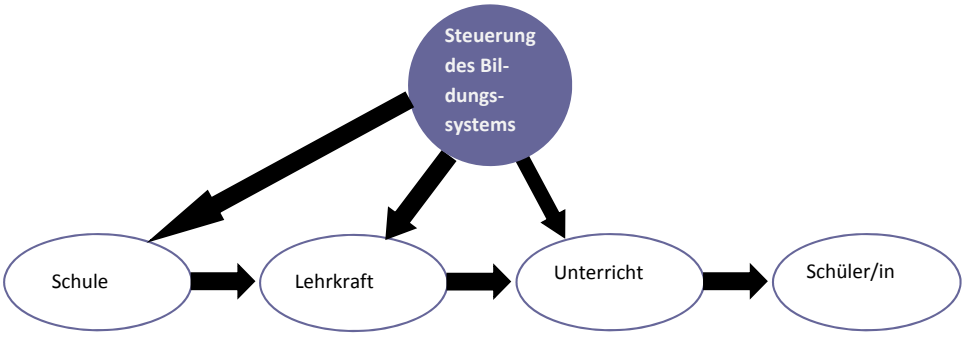


Abbildung 2: Einfluss des Systemmonitorings auf den Alltag einer Lehrkraft

Anders als beispielsweise in den USA gab es in Deutschland bis zum Anfang der 1990er Jahre keine Informationen darüber, wie erfolgreich schulische Bildungsprozesse in Deutschland verlaufen. Bildungsmonitoring fand bis zu diesem Zeitpunkt nicht oder nur unzureichend statt. Wie Bildungsmonitoring funktioniert, war bis zu diesem Zeitpunkt weder der Öffentlichkeit noch großen Teilen der Wissenschaft bekannt (Köller, 2008). Stattdessen fand in Deutschland bis vor einigen Jahren eine „reine Input-Steuerung" statt, der Output dagegen wurde nicht betrachtet.

Einstellung zu Bildungsmonitoring in Deutschland

Standardisierten Leistungsmessungen stand man in Deutschland lange Zeit kritisch gegenüber, sie wurden als „Übergriff der Bildungsverwaltung, als Angriff auf die professionelle Verantwortung der Lehrkräfte empfunden" (Klieme u.a., 2007). Zudem ging man davon aus, dass vergleichbare Klassen auch vergleichbare Lernfortschritte machen und deutsche Schülerinnen und Schüler im internationalen Vergleich gut abschneiden (KMK, 2006). Zwar wurde in der Wissenschaft viel über Qualitätsentwicklung diskutiert, die Qualitätssicherung dagegen wurde nicht angegangen (Klieme/Baumert, 2001).

Den Ausschlag dafür, dass das Bildungssystem intensiv und regelmäßig untersucht wird, gaben die Ergebnisse der Studie TIMSS im Jahr 1995. Durch die Rezeption der Ergebnisse wurden nicht nur inhaltliche Probleme des Bildungssystems deutlich, sondern auch, in welchem Rahmen sich Veränderungen ergeben könnten (Specht, 2008). Um Probleme des Schulsystems schneller erkennen und beheben zu können, finden Erhebungen im Rahmen des Bildungsmonitoring seitdem regelmäßig statt (vgl. Übersicht in Kapitel 3.3.6).

Konstanzer Beschlüsse

Durch die erste Teilnahme an internationalen Schulleistungsuntersuchungen wurde den deutschen bildungspolitischen Akteuren immer deutlicher, wie Bildungsmonitoring wirken kann und wo die Vorteile liegen. Daher entschied sich die Ständige Konferenz der Kultusminister der Länder (KMK) im Oktober 1997 in ihren ‚Konstanzer Beschlüssen' dazu, regelmäßiges Bildungsmonitoring in Deutschland zu implementieren. Von diesem Zeitpunkt an wurde die Qualitätssicherung im Bildungswesen zum zentralen Thema erklärt. Dazu gehörte die Teilnahme an internationalen Schulleistungsstudien und die Einführung von Bildungsmonitoring auf nationaler Ebene (PISA-Konsortium Deutschland, 2007). Zu den Aktivitäten auf nationaler Ebene zählt beispielsweise eine regelmäßige Bildungsberichterstattung (vgl. Kapitel 3.5.3.8).

Aufbauend auf diesem Beschluss und den folgenden Ergebnissen der Systemmonitoring-Studien entwickelte die KMK im Jahr 2002 sieben Handlungsfelder und setzte im Jahr 2006 nach weiteren Studien neue Schwerpunkte. Diese Handlungsfelder und Schwerpunkte sollen als erste Handlungen auf Grund des Bildungsmonitorings verstanden und daher in Kapitel 3.3.5 ‚Praktische Implikationen' näher betrachtet werden.

3.1.2 Ziele von Systemmonitoring-Studien

Systemmonitoring-Studien verfolgen gemeinsam mit einer kontinuierlichen Bildungsberichterstattung und einer verstärkten Förderung der Bildungsforschung das Ziel zu erfahren, ob und wo das Bildungssystem optimiert werden muss (Stanat, 2008). Das Ziel der Studien ist es, „auf einer breiten empirischen Basis die Beschreibung und Analyse der Erträge fachlichen Lernens in den Mittelpunkt" zu rücken (Köller, 2008). Auf Grund der Durchführung und der Interpretation der Studien ergibt sich eine erhöhte Transparenz bildungspolitischer Entscheidungsprozesse, dadurch kann Handlungsbedarf deutlich gemacht werden (Stanat, 2008). Zudem kann überprüft werden, ob Reformmaßnahmen, die auf Grund von Ergebnissen vorangegangener Studien umgesetzt wurden, ihre Wirkung entfalten können (Klieme u.a., 2007).

Betrachtung der Stärken und Schwächen der Schülerinnen und Schüler

Die Studien dienen auch dazu, empirisch abgesicherte Befunde darüber zu erhalten, in welchen (Kompetenz)-Bereichen die Schülerinnen und Schüler Stärken und Schwächen haben. Eine solche Information ist wichtig, damit „die Bildungschancen aller gewahrt werden" können und eine erhöhte Betrachtung der individuellen Voraussetzungen der Schülerinnen und Schüler stattfinden kann (KMK, 2005). Einige Studien untersuchen auch die Einstellungen und Motive der Schülerinnen und Schüler. Dadurch ergeben sich Informationen, die, gemeinsam mit den Informationen über die Kompetenzen, bei der Entscheidung über Schulstrukturen, Lehrpläne, die Lehrerbildung und die Schulbuchgestaltung von großer Relevanz sein können (Klieme u.a., 2001).

Abgesehen von diesen übergeordneten Zielen verfolgen die einzelnen Studien aber auch eigene Ziele. Als Beispiel sollen hier die Ziele der DESI-Studie zitiert werden:

„Die DESI-Studie soll erstmals aussagefähige Daten zu den Fähigkeiten deutscher Schülerinnen und Schüler im aktiven und passiven Gebrauch des Deutschen und Englischen erheben, dokumentieren, Erklärungsmodelle unter Einschluss personaler, unterrichtlicher und schulischer Faktoren ableiten und Optimierungsansätze für den Unterricht aufzeigen." (Eichler, 2003)

Ein weiteres Ziel von Systemmonitoring-Studien, das vor allem in der Öffentlichkeit dankbar aufgenommen wird, ist die Bereitstellung von so genannten ‚benchmarks' (Klieme u.a., 2001).

Definition: benchmarks

Der Begriff ‚benchmarks' beschreibt in diesem Kontext Leistungsergebnisse sehr erfolgreicher Bildungssysteme im In- und Ausland. Mit diesen Leistungsergebnissen kann dann das ‚eigene' Ergebnis verglichen werden. Dabei muss im Rahmen des Systemmonitorings darauf geachtet werden, dass die Ergebnisse erst durch Zusammenhanganalysen erklärt und anschließend Stärken und Schwächen eines Systems benannt werden können. Zu beachten ist, dass es im Systemmonitoring nicht ausschließlich um ein reines ‚benchmarking', geschweige denn um ein ‚ranking' der teilnehmenden Bildungssysteme geht (Klieme u.a., 2001) (siehe weitere Ziele in Kapitel 1.2).

Auch in Zukunft wird sich Deutschland dem internationalen Vergleich stellen, damit die „erforderliche Anschlussfähigkeit" gesichert werden kann. Zudem können so die entwickelten Bildungsstandards (s. Kapitel 3.5.2.1) an internationale Maßstäbe angelehnt und überprüft werden (KMK, 2006).

3.1.3 Vorstellung von verschiedenen Systemmonitoring-Studien

Viele Aspekte haben alle Studien des Systemmonitorings gemeinsam. Diese sollen hier zunächst erläutert werden. Im Anschluss daran besteht die Möglichkeit, sich über die einzelnen Studien näher zu informieren.

- Sie testen international eine bestimmte Kompetenz von Schülerinnen und Schülern einer festgelegten Altersgruppe.
- Die Untersuchung findet zeitgleich und mit denselben Aufgaben in allen teilnehmenden Staaten statt.
- Die Federführung liegt bei einer allgemein anerkannten Institution, die die Durchführung und Auswertung der Daten plant und begleitet.
- In einigen Teilnehmerstaaten werden zusätzliche Kompetenzen innerhalb der Studien getestet um einen intranationalen Vergleich zu ermöglichen.
- Die Federführung dieser Zusatz- oder Ergänzungsstudien liegt im jeweiligen Land selbst.

Keine Antworten von Systemmonitoring-Projekten

In Hinblick auf die Ergebnisse der Studien ist zu beachten: Studien wie PISA geben keine Antwort darauf, an welcher Stelle und auf welche Weise die gemessenen Kompetenzen erworben wurden und auch nicht,

wie und wo die gemessenen Unterschiede in den Kompetenzen entstanden sind. Dafür fehlen notwendige Kontrollgruppen und die Studien sind nicht als Längsschnittstudien angelegt. Allerdings kann durch Systemmonitoring-Studien deutlich werden, ob und wo es Zusammenhänge zwischen sozialer Herkunft und Bildungserfolg gibt. Dies wird durch den „systemvergleichenden Charakter der internationalen Leistungsstudien" möglich (Specht, 2008).

Definition: Längsschnittstudien

Längsschnittstudien beschreiben Studien, die eine Situation oder Kompetenz nicht nur zu einem Zeitpunkt untersuchen. In Längsschnittstudien werden die Testpersonen zu mindestens zwei Zeitpunkten getestet, sodass sich eine zeitliche Entwicklung darstellen lässt. In Deutschland werden Längsschnittstudien im Rahmen des Nationalen Bildungspanels (NEPS) (siehe *http://www.uni-bamberg.de/neps/*) durchgeführt.

Im Folgenden sollen die vier Studien TIMSS, PISA, IGLU und DESI kurz zusammengefasst vorgestellt werden. Es handelt sich nicht um eine komplette Darstellung der einzelnen Studien. Für einen Überblick über bereits durchgeführte und noch anstehende Studien bietet Kapitel 3.3.6 zwei Tabellen, die diese darstellen.

3.1.3.1 TIMSS (optional)[3]

TIMSS stand zunächst für ‚Third International Mathematics and Science Study' und ist die Nachfolgeuntersuchung zu FIMS und SIMS (First and Second International Mathematics Study) und FISS und SISS (First and Second International Science Study), die allerdings ohne deutsche Beteiligung durchgeführt wurden. Die Zielsetzung von TIMSS liegt darin, die Mathematik- und Naturwissenschaftsleistungen in der Grundschule (TIMSS I), in der Sekundarstufe I (TIMSS II) und in der Sekundarstufe II (TIMSS III[4]) gleichzeitig zu untersuchen. Deutschland beteiligte sich 1995 nur an TIMSS II und III, 2007 nur an TIMSS I.

Die Studie wurde ab 1999 mit dem Namen ‚Trends in International Mathematics and Science Study' alle vier Jahre durchgeführt, wobei sich Deutschland im weiteren Verlauf nur noch 2007 beteiligte.

Ergänzungsstudie

Im Jahr 1995 wurde zusätzlich n die Studie ‚TIMSS-Deutschland' durchgeführt. In dieser Ergänzungsstudie wurden Schülerinnen und Schüler aus den Klassen 7 und 8 zusätzlich in einer Längsschnittstudie untersucht. Hier wurde ein Schwerpunkt auf die motivationale Entwicklung und die verwendeten Unterrichtsmethoden gelegt. 2007 fand eine solche deutsche Ergänzungsstudie nicht statt.

Teilnehmende

Tabelle 2 zeigt eine Übersicht über die teilnehmenden Schülerzahlen in Deutschland:

	TIMSS I	TIMSS II	TIMSS III
1995	-----	7.000 Schülerinnen und Schüler	5.000 Schülerinnen und Schüler
2007	5.496 Schülerinnen und Schüler	----	----

Tabelle 2: Deutsche Teilnehmerzahlen TIMSS 1995 und 2007

International nahmen an TIMSS 2007 425.000 Schülerinnen und Schüler aus 59 Staaten teil.

TIMSS ist aber nicht eine reine Untersuchung der Kompetenzen der Schülerinnen und Schüler, sondern beschäftigt sich, beispielsweise im Jahr 1995, mit fünf Bereichen:

- International vergleichende Analysen von Lehrplänen und Lehrbüchern
- Schulleistungsuntersuchungen in den mathematisch-naturwissenschaftlichen Fächern
- Befragungen der Schulleiter und Fachlehrer
- Videoaufnahmen im Mathematikunterricht in Deutschland, Japan und den USA
- Ethnographische Fallstudien in Deutschland, Japan und den USA

3 Optionale Kapitel dienen der Vertiefung und können bei Bedarf genutzt werden.

4 Beachtet werden muss dabei, dass in Deutschland in TIMSS III auch das berufliche Schul- und Ausbildungswesen mit betrachtet wurde, sodass auch Auszubildende in Teilzeitberufsschulen Teil der getesteten Schülerinnen und Schüler waren. Auch in den anderen teilnehmenden Staaten wurden in TIMSS III jeweils unterschiedliche Bildungsgänge untersucht.

Inhaltlich wurden die drei Bereiche Mathematik (Arithmetik, Geometrie/Messen, Daten), Naturwissenschaften (Biologie, Physik, Geographie) und kognitive Anforderungsbereiche (Reproduzieren, Anwenden, Problemlösen) untersucht (Baumert / Lehmann, 1997; Baumert / Bos / Watermann, 1998; Bos / Bonsen / Baumert / Prenzel / Selter / Walther, 2008).

3.1.3.2 PISA (optional)

PISA ist mit Sicherheit die bekannteste Studie des Systemmonitorings. Die Abkürzung steht für ‚Programme for International Student Assessment'. PISA ist Teil des Indikatorenprogramms der OECD (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung), „dessen Ziel es ist, den OECD-Mitgliedsstaaten vergleichende Daten über die Ressourcenausstattung, die individuelle Nutzung sowie die Funktions- und Leistungsfähigkeit ihrer Bildungssysteme zur Verfügung zu stellen". Zudem werden auch Zusammenhänge zwischen den betrachteten Kompetenzen und der sozialen und kulturellen Herkunft der Schülerinnen und Schüler weltweit betrachtet.

Untersuchungsgegenstand

Untersucht werden in PISA Grundbildungsaspekte in den drei Bereichen Lesekompetenz, mathematische Grundbildung und naturwissenschaftliche Grundbildung. Dabei wird immer in jedem Jahr einer der drei Aspekte als Schwerpunkt betrachtet. Tabelle 3 gibt eine Übersicht über die Schwerpunkte.

	Lesekompetenz	Mathematische Grundbildung	Naturwissenschaftliche Grundbildung
2000	X		
2003		X	
2006			X
2009	X		
2011		X	
2014			X

Tabelle 3: Schwerpunktbereiche in PISA

‚Grundbildungsaspekte' bedeutet dabei, dass in PISA Kompetenzen erfasst werden, die „in modernen Gesellschaften für eine befriedigende Lebensführung in persönlicher und wirtschaftlicher Hinsicht sowie für eine aktive Teilnahme am gesellschaftlichen Leben notwendig" sind.

Beispiel: Aufgabe aus PISA 2000 – PIZZA

Eine Pizzeria bietet zwei runde Pizzas mit derselben Dicke in verschiedenen Größen an. Die kleinere hat einen Durchmesser von 30 cm und kostet 30 Zeds. Die größere hat einen Durchmesser von 40 cm und kostet 40 Zeds.

Beispielaufgabe 1: Mathematische Grundbildung

Bei welcher Pizza bekommt man mehr für sein Geld? Gib eine Begründung an.[5]

Es wird deutlich, dass durch Systemmonitoring-Studien nicht sinnfreies Wissen abgefragt wird, sondern das Wissen immer mit der Anwendung in der Praxis verknüpft wird.

Als Beispiel dafür sollen die konkreten Fragestellungen von PISA 2000 dargestellt werden:

Beispiel: PISA 2000: „Wie gut können Schülerinnen und Schüler …

… geschriebene Texte unterschiedlicher Art in ihren Aussagen, Absichten und ihrer formalen Struktur verstehen, einordnen und sachgerecht nutzen?"

… verständnisvoll mit Mathematik umgehen und mathematische Werkzeuge in einer Vielfalt von Kontexten einsetzen?"

… naturwissenschaftliches Wissen anwenden und Schlussfolgerungen ziehen, welche die natürliche Welt und durch menschliches Handeln in ihr vorgenommene Veränderung betreffen?"

… fächerübergreifende Basiskompetenzen (selbstreguliertes Lernen und Vertrautheit mit Computern) einsetzen?"

5 Weitere PISA-Beispielaufgaben werden auf http://pisa.ipn.uni-kiel.de/fr_reload.html?beispielaufgaben.html gesammelt vorgestellt.

Dabei wird vor allem beim letzten Punkt deutlich, dass neben den fachspezifischen Kompetenzen auch fachübergreifende Kompetenzen (so genannte ‚cross-curricular competencies') wie der Umgang mit dem Computer untersucht wurden.

Auf internationaler Ebene wurden zudem biografische Hintergründe und Familien- und Lebensverhältnisse der Schülerinnen und Schüler erfasst, um unterstützende Informationen für die Suche nach Begründungen der Ergebnisse zu erhalten.

Teilnehmende

Tabelle 4 gibt eine Übersicht über die teilnehmenden Staaten und die Anzahl der teilnehmenden Schülerinnen und Schüler:

	International: Schülerinnen und Schüler	Teilnehmende Staaten	Deutsche Schülerinnen und Schüler
2000	180.000	32	5.000
2003	250.000	41	4.660
2006	400.000	56	4.891
2009	470.000	65	4.979

Tabelle 4: Übersicht über an PISA teilnehmende Schülerinnen, Schüler und Staaten

Ergänzungsstudie

Die internationale Studie wurde um eine nationale Untersuchungen unter dem Namen PISA-E (‚E' für Ergänzung) erweitert. Eine solche Erweiterung ist in allen Systemmonitoring-Studien problemlos möglich, so lange es keine Reibungspunkte mit der internationalen Untersuchung gibt. Diese könnten beispielsweise darin liegen, dass die Ergänzungsstudien zu viel Zeit und Aufwand einnehmen und die teilnehmenden Schülerinnen und Schüler nicht mehr genügend Zeit und Konzentration für die internationale Studie haben.

Die Ergänzungsstudien gehen mit einer Vergrößerung der Stichprobe einher. Dieses Vorgehen ist notwendig, damit „aussagekräftige regionale Vergleiche innerhalb der Länder auf der Basis einer ausreichend großen Stichprobe" ermöglicht werden können (van Ackeren, 2006).

Durch eine nationale Untersuchung können einerseits sowohl die einzelnen Bundesländer als auch die Schulformen untereinander verglichen werden.

Ziel der Ergänzungsstudie

Ziel von PISA-E ist es, Stärken und Schwächen der Schülerinnen und Schüler in den Bundesländern herauszuarbeiten. Durch diesen Vergleich können die Leistungen der einzelnen Bundesländer in Beziehung zum deutschen Durchschnitt, zum OECD-Durchschnitt und auch zueinander gesetzt und analysiert werden. Dabei soll vermieden werden, dass PISA-E zu einer ‚Bildungsolympiade' wird, es geht nicht um Gewinner und Verlierer, sondern um den positiven Vergleich und die Übernahme erfolgreicher Konzepte (vgl. den Begriff ‚benchmarks' in Kapitel 3.1.1).

Zudem wurden noch weitere Untersuchungen im Rahmen von PISA durchgeführt. Als Beispiel soll hier PISA 2003 dienen. Hier wurde unter dem Namen PISA-I-Plus (wobei das ‚I' für die internationale Untersuchung steht) eine Längsschnittstudie durchgeführt. Dafür wurden weitere Schülerinnen und Schüler ausgewählt und über den Zeitraum eines gesamten Schuljahres untersucht.

PISA aktuell

Die letzte Erhebung von PISA fand im Sommer 2009 statt. An PISA 2009 nahmen international rund 470.000 Schülerinnen und Schüler aus 65 Staaten (mit allen OECD-Staaten) teil. Im Vergleich zur ersten Erhebung im Jahr 2000 mit 43 Teilnehmerstaaten, sowie zu den Erhebungen in den Jahren 2003 mit 41 Teilnehmerstaaten, und 2006 mit 57 Teilnehmerstaaten lässt sich eine deutliche Steigerung erkennen. (Deutsches PISA-Konsortium 2001; PISA-Konsortium Deutschland 2005, 2006, 2007, 2008, 2010; Baumert u.a. 2002; Kiper/Kattmann 2003; Kiper 2003a)

3.1.3.3 IGLU (optional)

Nach den Ergebnissen von TIMSS 1995 wurde deutlich, dass eine Beteiligung Deutschlands im Grundschulbereich hilfreich gewesen wäre, da so hätte festgestellt werden können, welche der in TIMSS aufgetretenen Defizite der Schülerinnen und Schüler bereits aus der Grundschulzeit resultieren. Daher kam die Möglichkeit gelegen, sich an IGLU zu beteiligen. IGLU steht für ‚Internationale Grundschul-Lese-Untersuchung' und untersucht international die Lesekompetenz von Grundschülerinnen und Grundschülern. Sie fand zum ersten Mal im Jahr 2001 statt. Im internationalen Rahmen wird die Studie ‚Progress in International Reading Literacy Study' (PIRLS) genannt. Geplant ist eine regelmäßige Durchführung alle fünf Jahre, sodass auch im Jahr 2006 Daten im Rahmen von IGLU erhoben wurden. Hier wurden „mittels

authentischer Texte[6] verschiedener Textgattungen unterschiedliche Aspekte der Kompetenz im Rahmen verschiedener Leseabsichten“ untersucht, zudem wurden „die Fähigkeit zum Schreiben und die kognitiven Lernvoraussetzungen ermittelt“ (Döbert et al., 2009). Die nächste Untersuchung findet im Jahr 2011 statt.

In Deutschland wurden 2001 zusätzlich zum Leseverständnis an einem zweiten Testtag auch die Bereiche Mathematik, Naturwissenschaften, Orthographie und Aufsatz im Rahmen der Studie IGLU-E (‚E‘ für Ergänzung) untersucht. In den Bereichen Mathematik und Naturwissenschaften konnten Test-Items aus TIMSS (s. Kapitel 3.1.3.1) verwendet werden, sodass die Ergebnisse der getesteten Schülerinnen und Schüler nachträglich auf der internationalen TIMSS-Skala eingeordnet werden konnten. Für den Bereich Orthographie wurde ein Lückentext verwendet, in dem Wörter diktiert wurden, die dem Grundwortschatz von Viertklässlern entsprechen sollten. 2001 beteiligten sich allerdings nur 12 der 16 Bundesländer an dieser zusätzlichen Untersuchung, im Jahr 2006 nahmen alle 16 Bundesländer teil.

Einordnung auf TIMSS-Skala

Auch in dieser Studie wurden Fragebögen an Eltern, Lehrer, Schulleiter und Schülerinnen und Schüler verteilt und somit viele Zusatzinformationen erhoben, um Informationen für eine Interpretation der Ergebnisse und einen Einblick in die Lebenswelt der Schülerinnen, Schüler, Lehrer, Eltern und Schulleiter zu erhalten.

Tabelle 5 gibt eine Übersicht über die an IGLU/PIRLS teilnehmenden Staaten und die Anzahl der Schülerinnen und Schüler:

	Internationale Schülerinnen und Schüler	Teilnehmende Staaten bzw. Regionen	Deutsche Schülerinnen und Schüler
2001	146.490	35	10.571[7]
2006	215.137	45	8.302[8]

Tabelle 5: Übersicht teilnehmende Staaten, Schülerinnen und Schüler an IGLU/PIRLS

Für die Studie wurden in allen Ländern Schülerinnen und Schüler aus der Klassenstufe getestet, die die meisten neunjährigen Kinder umfasste. In den meisten Teilnehmerstaaten, so auch in Deutschland, entspricht dies der Jahrgangsstufe 4. Ausgenommen waren aus der Zufallsauswahl solche Klassen, in denen die Kinder aufgrund geistiger oder körperlicher Behinderung nicht zur Bearbeitung der Aufgaben in der Lage waren (vgl. Kapitel 3.3.5).

(Bos u.a., 2003, 2007, 2008)

3.1.3.4 DESI (optional)

An DESI nahmen Schülerinnen und Schüler der Jahrgangsstufe 9 teil. Die Abkürzung DESI steht für ‚Deutsch Englisch Schülerleistungen International‘. Die Studie wurde im Jahr 2001 von der KMK in Auftrag gegeben, um die „sprachlichen Leistungen und die Unterrichtswirklichkeit in den Fächern Deutsch und Englisch“ zu untersuchen. Das ‚I‘ in DESI steht für die Teilnahme der Länder Schweiz und Österreich, zudem sollten Aspekte des Erwerbs der Fremdsprache Englisch international verglichen werden. Ziel von DESI war, die Leistungen der Schülerinnen und Schüler sowohl in der ersten Fremdsprache Englisch als auch in der Unterrichtssprache Deutsch in den Bereichen ‚aktiver Sprachgebrauch‘, ‚Sprachbewusstheit‘ und ‚Kommunikation‘ zu untersuchen (vgl. Beschreibung des Ziels von DESI in Kapitel 3.1.2). Dabei wurden schriftliche und mündliche Kompetenzen und somit alle Teilkompetenzen und alle Lernbereiche des Unterrichts erfasst.

Untersuchungsgegenstand

Der Begriff der ‚Unterrichtssprache‘ wird hier gewählt, da die Studie davon ausgeht, dass ein großer Anteil der Schülerinnen und Schüler Deutsch nicht als Erstsprache spricht. Daher wurden alle Schülerinnen und Schüler in der Auswertung danach kategorisiert, ob sie ‚Deutsch als Erstsprache‘ oder ‚Deutsch als Zweitsprache‘ sprechen oder ob sie bilingual aufgewachsen sind.

Unterrichtssprache

DESI kann somit als Ergänzung zu PISA verstanden werden, da dort nur die Lesekompetenz der Schülerinnen und Schüler in der Unterrichtssprache getestet wurde.

Am Anfang und am Ende des Schuljahres 2003/2004 wurden in Deutschland 10.639 Schülerinnen und Schüler der neunten Klassen aller Schulformen getestet (insgesamt 219 Schulen, davon 40 mit bilingualem Zweig). Somit handelt es sich bei DESI um eine Längsschnittstudie (vgl. Definition in Kapitel 3.1.3).

Durchführung

6 Authentische Texte bezeichnen Texte, die in der Realität verwendet wurden oder theoretisch verwendet werden könnten.

7 Nur 7.633 Schülerinnen und Schüler erfüllten die Anforderungen des Zufallkriteriums, sodass nur diese Anzahl in die Auswertung eingegangen ist.

8 Einige Schülerinnen und Schüler waren am Testtag nicht anwesend, andere hatten keine Genehmigung der Eltern, sodass letztendlich 7.899 Schülerinnen und Schüler aus Deutschland teilgenommen haben.

Zusätzlich zu den Schülerinnen und Schülern wurden die Lehrer, Eltern und Schulleitungen befragt und es wurden Videoaufnahmen im Englischunterricht durchgeführt. Durch die Anlage der Studie konnten bundesweit repräsentative Ergebnisse erzielt und durch die zusätzlichen Befragungen Erkenntnisse über Lehr- und Lernprozesse und den sprachlichen Kompetenzerwerb gewonnen werden. Diese Informationen sind sowohl für das allgemeine Systemmonitoring als auch für die Praxis von großer Bedeutung.

getestete Kompetenzen

Die von DESI getesteten Teilkompetenzen werden in der folgenden Tabelle dargestellt:

	Deutsch	Englisch
Hörverstehen		X
Leseverständnis	X (auch: Lesegeschwindigkeit)	X (global)
Wortschatz	X	
Grammatik	X	
Sprachbewusstheit	X (Grammatik, Stil und Rechtschreibung)	X (Grammatik und Adressatengerechtigkeit)
Mündliche Sprachkompetenz	X (Kommunikation/ Argumentation)	X
Rechtschreibung (Satzkonstruktion, Orthographie, Zeichensetzung)	X	
Schreiben	X	X (kreativ)
Interkulturelle Kompetenz		X
Globale Sprachkompetenz		X

Tabelle 6: DESI-Teilkompetenzen in Deutsch und Englisch

Auch wenn sich diese Studie in einigen Bereichen deutlich von den anderen Studien des Systemmonitorings unterscheidet, ist ihre Betrachtung wichtig, da die Ergebnisse (vgl. Kapitel 3.4.2.4) gezeigt haben, dass „eine zentrale Ursache für die unbefriedigenden Ergebnisse im Unterrichtsgeschehen selbst zu suchen ist“ und als Konsequenz daraus mit der Einführung der Bildungsstandards (vgl. Kapitel 3.5.3.1) versucht wurde, dieser Schwäche entgegenzuwirken (Köller, 2008).

(Klieme, 2006; Klieme/Baumert, 2001; Eichler, 2003, Beck/Klieme, 2007; DESI-Konsortium, 2008)

3.1.4 Weiterführende Literatur

Eine Übersicht zum Thema ‚Bildungsmonitoring‘ bietet die Dokumentation zum OECD/CERI-Regionalseminar für die deutschsprachigen Länder in Potsdam („Bildungsmonitoring, Vergleichsstudien und Innovation“, Hg: LISUM, mn:ukk, EDK, 2008). Detaillierte Angaben über die Charakteristika der einzelnen Studien bieten jeweils die Veröffentlichungen der Studien. Zudem vermittelt das Buch ‚Leistungsmessungen in Schulen‘ von Franz E. Weinert (2001) sehr ausführliche Informationen über das Thema.

Beispielaufgaben können im Internet unter den folgenden Links gefunden werden:

PISA: http://pisa.ipn.uni-kiel.de/fr_reload.html?beispielaufgaben.html

TIMSS: http://timss.ifs-dortmund.de/31.html

3.1.5 Verständnis-Aufgaben und Diskussionspunkte

1. *Eine Schule hat an PISA teilgenommen und ein Ergebnis ähnlich dem finnischen erreicht. Der Schulleiter lädt zu einer Pressekonferenz mit der Überschrift: „Wir sind besser als Finnland“. Welche Fragen könnten die Journalisten ihm stellen?*
2. *Hinterfragen Sie die Entscheidung, bei PISA ein bestimmtes Alter und nicht eine Jahrgangsstufe untersucht zu haben.*
3. *In welchen Bereichen des Studiendesigns unterscheidet sich die PISA-Studie von der DESI-Studie und welche Bedeutung hat das für die Interpretation der Ergebnisse?*
4. *Formulieren Sie Gründe, warum ein Land nicht an einer (Teil-)Erhebung teilnehmen möchte/teilnimmt/teilnehmen kann und suchen Sie Argumente dafür, dass sich immer mehr Staaten im Laufe der Zeit dazu entschieden haben, beispielsweise an PISA teilzunehmen.*

3.2 Bewertungskriterien im Systemmonitoring

Kapitel 3.2 beschreibt die drei Bezugsnormen sozial, kriterial und individuell. Diese Bezugsnormen werden ebenfalls in den Studienbrief-Teilen ‚Individuelle Diagnostik' und ‚Vergleichsarbeiten' beschrieben, in Kapitel 3.2 allerdings mit Bezug auf das Systemmonitoring und die damit verbundenen Studien. Zudem wird auf die Unterscheidung zwischen einer curricularen und einer ‚literacy'-Orientierung eingegangen. Es schließen sich Ausführungen über Kompetenzen, Kompetenzmodelle und Kompetenzniveaus und deren Zusammenhang mit der Standardentwicklung an.

In diesem Kapitel werden folgende Fragen beantwortet:

- *Wie werden die drei Bezugsnormen (kriterial, sozial, individuell) in den Systemmonitoring-Studien angewendet?*
- *Was bezeichnet der Begriff ‚literacy'?*
- *Worin unterscheidet sich eine ‚literacy'-Orientierung von einer curricularen Orientierung?*
- *Was sind Kompetenzmodelle und Kompetenzniveaus, wie werden sie erstellt und wie funktionieren sie?*
- *Was sind Standards und wie hängen sie mit Kompetenzmodellen und Kompetenzniveaus zusammen?*

3.2.1 Beschreibung der Bezugsnormen

Für Leistungsvergleiche werden Maßstäbe beziehungsweise ein Referenzrahmen benötigt. Drei mögliche Bezugsnormen kommen in Frage: ‚kriterial', ‚sozial' und ‚individuell'. Da es sich bei den Studien des Systemmonitorings um Large-Scale Assessments (vgl. Definition in Kapitel 3.1) handelt, werden nicht Testergebnisse einzelner Schülerinnen und Schüler sondern Testergebnisse einzelner Staaten miteinander verglichen.

3.2.1.1 Kriteriale Bezugsnorm

Kriteriale Bezugsnorm bedeutet, dass ein objektiv definiertes Kriterium als Vergleichsmaßstab dient und eine bestimmte Kompetenz erreicht werden muss (Näheres zum Begriff Kompetenz in Kapitel 3.2.3.1 und in Studienbrief-Teil Vergleichsarbeiten).

Die benötigten Kriterien können auf unterschiedliche Art gewonnen werden. Eine Möglichkeit ist der Rückgriff auf die Lehrpläne. Bei einer internationalen Studie ist dazu die Erarbeitung eines ‚weltweiten Curriculums' notwendig, so ist beispielsweise bei TIMSS vorgegangen worden. Die Alternative ist eine Orientierung an Grundkompetenzen, auf die sich die beteiligten Wissenschaftlerinnen und Wissenschaftler verständigen.

Die Orientierung an einer kriterialen Bezugsnormen ermöglicht außerdem eine differenzierte Auswertung, die über die reine Feststellung von Mittelwerten hinausgeht, in dem z.B. angegeben werden kann, wie viele Schülerinnen und Schüler welches der als Kriterium definierten Kompetenzniveaus erreicht haben.

Da aber eine bloße Aussage über das (Nicht-)Erreichen bestimmter Kompetenzen beziehungsweise Standards nicht alle Ziele der Systemmonitoring-Studien abdecken kann (vgl. die Ziele der Studien in Kapitel 3.1.2), muss auch die soziale Bezugsnorm betrachtet werden. Insbesondere TIMSS und PISA zeigen, wie eine solche Verbindung funktionieren kann (Klieme u.a., 2007).

3.2.1.2 Soziale Bezugsnorm

Eine soziale Bezugsnorm besteht, wenn ein Ergebnis mit den Ergebnissen einer Referenzgruppe verglichen wird. In den Systemmonitoring-Studien existiert diese soziale Bezugsnorm, da die Teilnehmerstaaten und in den Ergänzungsstudien auch die einzelnen Bundesländer verglichen werden. Somit kann diese Bezugsnorm im Rahmen des Bildungsmonitorings auch ‚nationale' beziehungsweise ‚internationale' Bezugsnorm genannt werden.

Vergleich mit ‚ähnlichen' Ländern

IGLU zeigt einen sinnvollen Umgang mit Vergleichen nach sozialen Bezugsnormen: Generell bietet die Studie die Möglichkeit, sich mit jedem teilnehmenden Staat zu vergleichen. Allerdings ist es sinnvoller, sich mit Staaten zu vergleichen, „die einen ähnlichen wirtschaftlichen und kulturellen Hintergrund haben". Nur so kann abgeschätzt werden, inwieweit die Leistungen der Schülerinnen und Schüler „als ein Ergebnis des Einsatzes von Ressourcen und kultureller Erfahrung zu betrachten und zu bewerten" sind (Bos u.a., 2003). Die folgende Tabelle stellt die ‚IGLU-Ländervergleichsgruppen' dar.

Vergleichsgruppe 1[9]	Vergleichsgruppe 2[10]		Vergleichsgruppe 3[11]
Deutschland	Bulgarien	Niederlande	Argentinien
England	Deutschland	Norwegen	Belize
Frankreich	England	Rumänien	Iran
Griechenland	Frankreich	Schottland	Kolumbien
Italien	Griechenland	Schweden	Kuweit
Niederlande	Island	Slowakei	Marokko
Schottland	Italien	Slowenien	Mazedonien
Schweden	Kanada	Tschechien	Türkei
	Lettland	Ungarn	
	Litauen	USA	
	Neuseeland	Zypern	

Tabelle 7: Ländervergleichsgruppen bei IGLU (Bos u.a., 2003)

Die drei Vergleichsgruppen dienen dazu, dass neben dem internationalen Mittelwert auch die Mittelwerte der drei Vergleichsgruppen als weitere Kennwerte für die Analyse angegeben werden können und so eine bessere Einschätzung über das eigene Abschneiden möglich ist. Es handelt sich also um eine Art ‚fairen Vergleich' (vgl. Studienbrief-Teil Vergleichsarbeiten).

In den Bildungsmonitoring-Studien ist sowohl die soziale als auch die kriteriale Bezugsnorm von großer Bedeutung.

3.2.1.3 Individuelle Bezugsnorm

Bei der ‚individuellen' oder ‚verlaufsorientierten' Bezugsnorm wird durch wiederholte Einsetzung eines Tests in der gleichen Probandengruppe die Veränderung der Kompetenzen betrachtet.

Bei den Untersuchungen des Systemmonitorings ist diese Betrachtungsweise auf Grund der jeweils neu gezogenen Stichprobe und der wechselnden Untersuchungsschwerpunkte nicht sinnvoll.

‚individuelle' Vergleichbarkeit der Länder

Allerdings kann eine individuelle Vergleichbarkeit dargestellt werden, wenn jedes Land als ein ‚Individuum' angesehen wird. Dadurch, dass die Studien in regelmäßigen Abständen durchgeführt werden, können jedoch beispielsweise die Ergebnisse von PISA 2000 und PISA 2003 innerhalb der einzelnen Staaten verglichen werden. Dabei ist aber zu beachten, dass bei PISA jeweils ein Schwerpunktthema verstärkt untersucht wird und somit ein intensiver Vergleich nur alle neun Jahre möglich ist. Dies funktioniert allerdings nur bei gleichbleibender Rahmenkonzeption. Über eine Veränderung dieser wird vor jeder Erhebung von den Mitgliedsstaaten entschieden. Aus einer Veränderung der Rahmenkonzeption folgt dann, dass bei der Betrachtung der Trends Einschränkungen entstehen können, da durch eine Konzeptionsveränderung teilweise andere Kompetenzen untersucht werden.

Durch einen Vergleich nach individueller Bezugsnorm werden Leistungssteigerungen oder Leistungsrückgänge der einzelnen Staaten deutlich und auch der Veränderungsbedarf im Bildungswesen kann hier, allerdings ‚nur' auf der Ebene eines Staates, dargestellt werden.

3.2.2 ‚Literacy'-Orientierung vs. curriculare Orientierung

‚literacy' (engl.) = Literalität (dt.)?

Ganz allgemein kann der Begriff ‚literacy' mit dem Begriff ‚Literalität' übersetzt werden. In der deutschen Literatur wird mit dem Begriff der ‚Literalität' häufig allerdings nur die Fähigkeit zu lesen und zu schreiben bezeichnet. Der vorliegende Studienbrief-Teil schließt sich einer weiter gefassten und in diesem Kapitel beschriebenen Definition an und verwendet im Folgenden den Begriff ‚literacy', ohne ihn zu übersetzen. Diese Variabilität des weit gefassten Begriffs der ‚literacy' wird auch dadurch deutlich, dass von ‚computer literacy', ‚science literacy' oder auch ‚health literacy' gesprochen werden kann (Grotlüschen/Linde, 2007), sodass die in ‚literacy' beinhalteten Kompetenzen weit über das reine Schreiben und Lesen hinausgehen, diese Kompetenzen aber weiterhin wichtiger Bestandteil der ‚literacy' sind.

Nach einer solchen Auffassung „gehören die Beherrschung der Muttersprache in Wort und Schrift sowie ein hinreichend sicherer Umgang mit mathematischen Symbolen und Modellen zum Kernbestand kultureller Literalität" (Köller, 2008). Dazu gehören „sprachliche Kompetenzen, die grundlegende Formen des kommunikativen Umgangs mit der Welt repräsentieren" (Klieme/Baumert/Köller, 2000).

9 In Vergleichsgruppe 1 befinden sich die Staaten der europäischen Union, die an IGLU teilgenommen haben.
10 Vergleichsgruppe 2 beinhaltet alle Teilnehmerstaaten der OECD ohne die Türkei, dafür alle Staaten, die bis 2007 Mitglied der europäischen Union wurden.
11 Vergleichsgruppe 3 klassifiziert sich nach den Ergebnissen in IGLU: Hier werden die Staaten aufgeführt, die beim Lesen mehr als eine halbe Standardabweichung unterhalb des internationalen Mittelwertes liegen.

Aber auch eine ‚computer literacy' gehört zu den Kulturwerkzeugen der heutigen Welt. All diese Kompetenzen dienen in Kombination dazu, dass Menschen in ihrem Alltag, im Beruf und in der Öffentlichkeit selbstverantwortlich handeln können (Klieme u.a., 2001). Defizite in diesen Bereichen der Grundbildung führen dazu, dass die Teilnahme an der allgemeinen gesellschaftlichen Entwicklung gefährdet ist und Perspektiven in den Bereichen ‚Beruf' und ‚Leben' schlechter sind (Köller, 2008).

Defizite

Bestimmte Bereiche der Grundbildung sind aber ‚wichtiger' als andere. In Bezug auf ‚scientific literacy' heißt das, dass die meisten Menschen mit einem naturwissenschaftlichen Alltagswissen auskommen, naturwissenschaftliches Spezialwissen dagegen nicht beherrscht werden muss. Hier unterscheidet man zwischen ‚scientific literacy', die nur von etwa 10 Prozent eines Jahrgangs erreicht wird, und ‚functional literacy', d.h. Alltagswissen, über das etwa 30 Prozent eines Jahrgangs verfügen (Baumert/Lehmann, 1997).

‚wichtigere' Bereiche

Vertiefung: Anpassung des ‚literacy'-Konzepts in PISA (optional)

Für jeden Bereich der ‚Grundbildung' wird das ‚literacy'-Konzept entsprechend angepasst (Kiper, 2003b). Als Beispiel werden die Definitionen von ‚scientific literacy', ‚reading literacy' und ‚mathematical literacy' dargestellt, wie sie in PISA verwendet werden.

Definition: ‚scientific literacy' in PISA

„Naturwissenschaftliches Wissen anzuwenden, naturwissenschaftliche Fragen zu erkennen und aus Belegen Schlussfolgerungen zu ziehen, um Entscheidungen zu verstehen und zu treffen, die die natürliche Welt und die durch menschliches Handeln an ihr vorgenommenen Veränderungen betreffen." (PISA-Konsortium, 2001)

Definition: ‚reading literacy' in PISA

„[…] die Fähigkeit, geschriebene Texte unterschiedlicher Art in ihren Aussagen, ihren Absichten und ihrer formalen Struktur zu verstehen und sie in einen größeren sinnstiftenden Zusammenhang einzuordnen, sowie in der Lage zu sein, Texte für verschiedene Zwecke sachgerecht zu nutzen." (PISA-Konsortium, 2001)

Definition: ‚mathematical literacy' in PISA

„[…] ein Verständnis der Rolle, die Mathematik in der sozialen, kulturellen und technischen Welt spielt, und die Fähigkeit, Sachverhalte unter mathematischen Gesichtspunkten angemessen zu beurteilen: […] auch die Fähigkeit […] Mathematik aktiv zu nutzen, um Anforderungen des Alltags zu bewältigen." (PISA-Konsortium, 2001)

Innerhalb der ‚literacy'-Orientierung kann man mehrere Stufen unterscheiden, die „von einem anschaulichen Verständnis von Alltagsphänomenen über ein sinnvolles Anwenden elementarer Modelle bis hin zur vollen Kommunikations- und Urteilsfähigkeit in diesen Gebieten reichen" (Klieme u.a., 2001).

Vor Beginn einer Studie des Systemmonitorings können sich die Administratorenbeziehungsweise die Auftraggeber entscheiden, ob sie die im Curriculum beschriebenen Kompetenzen oder aber ‚literacy' als Grundlage untersuchen wollen, sodass sich eine ‚literacy'-Orientierung oder eine curriculare Orientierung ergibt. Bei beiden Orientierungen handelt es sich um Formen der kriterialen Bezugsnorm (vgl. Kapitel 3.2.1.1).

Als Beispiel soll TIMSS dienen, da hier beide Orientierungen in der Studie vertreten sind (Klieme u.a., 2001).

Beispiel: curriculare und ‚literacy'-Orientierung in TIMSS (1995)

Curriculare Orientierung:

TIMSS II und TIMSS III lehnten sich eng an die jeweiligen Lehrpläne an. Ziel war es hier, „zentrale Inhalte und Anforderungen des Fachunterrichts" (Klieme u.a., 2001) zu untersuchen. Dies ließ sich vor allem dadurch gut ermöglichen, dass die Curricula international relativ übereinstimmend waren.

‚literacy'-Orientierung:

Die mathematische und naturwissenschaftliche Grundbildung in TIMSS III wurde nach dem ‚literacy'-Konzept untersucht. Unter Grundbildung wurde hier „die Kenntnis zentraler Konzepte und Arbeitsprin-

zipien dieser Fächer“, aber auch „die Fähigkeit, dieses Wissen in alltäglichen Zusammenhängen zu nutzen und zu kommunizieren“ verstanden. Hier wird besonders deutlich, dass ‚literacy‘ „eine Verknüpfung von fachsystematischem Verständnis und Anwendungsorientierung“ beschreibt.

Dabei orientierte sich der TIMSS-Grundbildungstest an folgenden Merkmalen des ‚literacy‘-Konzepts (Klieme/Baumert/Köller, 2000):

- Betonung zentraler theoretischer Konzepte
- Einschränkung der stofflichen Breite zu Gunsten der Möglichkeit, in einzelnen Gebieten tieferes Verständnis zu erreichen
- Verstärkung fachübergreifender und fächerverbindender Ansätze
- Betonung des selbstständigen mathematischen und naturwissenschaftlichen Handelns und Kommunizierens

Vermischung der Orientierungen

Die Vermischung der beiden Orientierungen in TIMSS wurde allerdings stets kritisch diskutiert. Im Gegensatz dazu steht PISA: Hier wird die curriculare Orientierung in den Hintergrund gedrängt und stattdessen das ‚literacy‘-Konzept verfolgt (PISA-Konsortium, 2002). Das, was in PISA unter ‚Bildung‘ verstanden wird, „orientiert sich damit zunächst an der Funktion von Kompetenzen im kulturellen und gesellschaftlichen Zusammenhang“. Bei der Ausarbeitung der Grundkonzeption von PISA wurde hinterfragt, welches Konzeptwissen (‚Wissen, dass‘) und welches Prozesswissen (‚Wissen, wie‘) für die Teilhabe am gesellschaftlichen Leben wichtig ist und in welchen Situationen welches Wissen angewendet werden soll (PISA-Konsortium Deutschland, 2007).

Vertiefung: Konzept- und Prozesswissen (optional)

Konzeptwissen (Wissen, dass)		
Definition: Wissen über die Wechselbeziehungen zwischen den einzelnen Elementen des Basiswissens innerhalb eines größeren Zusammenhangs, das ein gemeinsames Funktionieren sichert		
Kenntnis der Klassifikationen und Kategorien (z. B. die verschiedenen geologischen Zeitperioden)	Kenntnis der Prinzipien und Verallgemeinerungen (z. B. Theoreme und Gesetze)	Kenntnis der Theorien, Modelle und Strukturen (z. B. Evolutionstheorie)

(nach: ThILLM – Thüringer Institut für Lehrerfortbildung, Lehrplanentwicklung und Medien; Gehirngerechtes Klassenzimmer – “Handreichungen für die Unterrichtspraxis”; ThILLM-Heft 126)

Prozesswissen (Wissen, wie)		
Definition: Wissen darüber, wie man etwas tut; Wissen über Methoden des Nachforschens; Kenntnis über Anwendungskriterien für Fähigkeiten, Algorithmen, Techniken und Methoden		
Kenntnis fachspezifischer Fähigkeiten und Algorithmen (z. B. zur Lösung einer quadratischen Gleichung)	Kenntnis fachspezifischer Techniken und Methoden (z. B. über die Interpretation eines literarischen Textes)	Kenntnis der Kriterien zur Anwendung bestimmter Verfahrensweisen (z. B. welche Methoden zu benutzen sind, um Informationen eines Textes zu visualisieren)

(nach: ThILLM – Thüringer Institut für Lehrerfortbildung, Lehrplanentwicklung und Medien; Gehirngerechtes Klassenzimmer – “Handreichungen für die Unterrichtspraxis”; ThILLM-Heft 126)

IGLU

Auch in IGLU orientiert sich das Konzept der Lesekompetenz an ‚literacy‘, sodass unter ‚reading literacy‘ hier die Kompetenz gemeint ist, „Lesen in unterschiedlichen, für die Lebensbewältigung praktisch bedeutsamen Verwendungssituationen einsetzen zu können“ (Bos u.a., 2003). Somit ergibt sich die theoretische Struktur der Lesekompetenz wie in Abbildung 3 dargestellt:

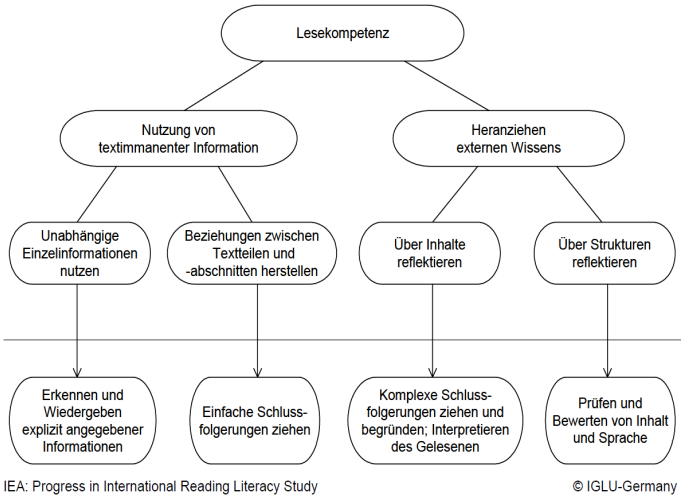


Abbildung 3: Theoretische Struktur der Lesekompetenz in IGLU

Die Abbildung 3 zeigt, dass die Lesekompetenz bei IGLU in zwei theoretische Dimensionen eingeteilt wird. Dies ist einerseits die Verstehensleistung von textimmanenten, andererseits die Verstehensleistung von wissensbasierten Informationen durch externes Wissen. Diese beiden Dimensionen werden dann in weitere Aspekte unterteilt, die sich durch die unterschiedlichen kognitiven Anforderungen ergeben (Bos u.a., 2003).

Lesen für den Alltag

Hierbei wird deutlich, dass Lesen nicht nur für die reine Wissensaneignung sondern vielmehr für den alltäglichen Gebrauch benutzt werden soll. Zudem soll sichergestellt werden, dass auch nach Abschluss der Schule die Lesekompetenz so gestaltet ist, dass die Schülerinnen und Schüler auf weiteres Lernen vorbereitet sind (Bos u.a., 2003).

DESI

Ein Beispiel für eine Studie mit curricularer Orientierung ist DESI (vgl. Kapitel 3.1.3.4). Bei DESI wurde vor Beginn der Testentwicklung eine detaillierte Lehrplananalyse durchgeführt, sodass die Aufgaben in Inhalt und im Anspruch eng an die Lehrpläne angelehnt sind (Klieme, 2006).

curriculare Vereinheitlichung

Möglich ist, dass sich durch die internationalen Studien des Systemmonitorings eine curriculare Vereinheitlichung in den OECD-Staaten entwickeln wird. In einem solchen Curriculum werden dann kulturelle und nationale Besonderheiten immer weniger bedeutsam. Das ‚literacy'-Konzept kann bei der Entwicklung eines solchen Curriculums behilflich sein, da dadurch den Studien bereits ein internationales Curriculum zur Verfügung steht. Somit erfolgt „eine Verschiebung in Richtung Wertschätzung anwendungsorientierten Wissens und handlungsbezogener Kompetenzen" (Kiper, 2003b). Was in diesem Zusammenhang genau unter ‚Kompetenzen' verstanden wird, wird im folgenden Kapitel erläutert.

3.2.3 Kompetenz, Kompetenzniveaus, Kompetenzmodelle und Standards

Das Kapitel beginnt mit einer Definition des Begriffs ‚Kompetenz'. Es folgen Ausführungen zu ‚Kompetenzmodellen' und ‚Kompetenzniveaus' und deren Verbindung mit Standards. Damit keine Doppelungen mit den beiden anderen Studienbrief-Teilen entstehen, wird darauf geachtet, dass alle Erläuterungen eine Verbindung zum Systemmonitoring aufweisen.

3.2.3.1 Kompetenz

Im Studienbrief „Vergleichsarbeiten" wird der Begriff ‚Kompetenz' eingehend definiert, daher soll an dieser Stelle darauf verwiesen werden und hier nur einige ergänzende Aspekte mit Bezug auf das Systemmonitoring präsentiert werden.

Die Aufgaben in den Studien des Systemmonitorings fragen kein reines Faktenwissen ab, sondern untersuchen die Möglichkeiten der Schülerinnen und Schüler, in bestimmten Kontexten Anforderungen zu erfüllen und Probleme zu lösen. Je nachdem wie die Schülerinnen und Schüler die Aufgaben beantworten, wird ihnen eine entsprechende Kompetenz zugesprochen (PISA-Konsortium, 2008).

Somit beschreibt ‚Kompetenz' die Fähigkeit, bestimmte Probleme und Anforderungen zu lösen. Das folgende Beispiel zeigt, wie eine Fremdsprachenkompetenz beschrieben werden kann, die diese Aspekte beachtet. Das zu erreichende Ziel beziehungsweise das zu lösende Problem besteht hier in der kommunikativen Handlungsfähigkeit.

Beispiel: Fremdsprachenkompetenz

(Bildungsziel: kommunikative Handlungsfähigkeit) drückt sich darin aus, …

- … wie gut man kommunikative Situationen bewältigt
- … wie gut man Texte unterschiedlicher Art versteht
- … dass man adressatengerecht Texte verfassen kann
- … dass man grammatische Strukturen korrekt aufbauen und bei Bedarf korrigieren kann
- … dass man sich in der Intention und Motivation offen und akzeptierend mit anderen Kulturen auseinander setzen kann (Klieme, 2007).

Kompetenzen abgrenzbar

Von besonderer Bedeutung ist es, dass Kompetenzen voneinander abgrenzbar sind. Das heißt, dass eindeutig bestimmt werden kann, ob eine Schülerin oder ein Schüler eine bestimmte Kompetenz besitzt oder nicht besitzt. Daher werden die Kompetenzen sehr konkret beschrieben, sodass sie leicht in entsprechende Aufgaben umgesetzt und getestet werden können (KMK, 2005).

Vorteile

Dass im Unterricht und in den Studien zum Systemmonitoring Kompetenzen untersucht werden, hat die folgenden Vorteile (KMK, 2005):

- Der Blick wird auf die Lernergebnisse der Schülerinnen und Schüler gelenkt.
- Das Lernen wird auf die Bewältigung von Anforderungen und nicht nur auf den Aufbau von – zunächst – ungenutztem Wissen ausgerichtet.
- Das Lernen wird als kumulativer Prozess organisiert.

3.2.3.2 Kompetenzmodelle

Die Aufgaben und Charakteristika von Kompetenzmodellen sind:

Aufgabe der Modelle

- zu beschreiben, welche Lernergebnisse von Schülerinnen und Schülern zu bestimmten Zeitpunkten in den einzelnen Fächern erwartet werden
- aufzuzeigen, auf welche Art und Weise Wissen und Können erreicht werden können
- dabei zu helfen, zwischen abstrakt formulierten Bildungszielen und Aufgabensammlungen zu vermitteln
- die Basis dafür zu legen, dass Bildungsziele operationalisiert (d.h. messbar gemacht werden) und diese durch empirische Tests überprüft werden können
- nicht nur bei der Herstellung von Testverfahren, sondern auch in der Unterrichtspraxis zu helfen
- Anhaltspunkte zu bieten, sodass sich der Unterricht an den „Lernprozessen und Lernergebnissen der Schülerinnen und Schüler im jeweiligen Lernbereich“ orientieren kann (Klieme, 2007).

Entwicklung der Modelle

Die Kompetenzmodelle werden von internationalen Expertengruppen auf Basis der „pädagogischen und fachdidaktischen Forschung“ entwickelt (Klieme, 2007). Eine Anforderung an Kompetenzmodelle ist, sowohl im oberen, mittleren, als auch im unteren Leistungsbereich auszudifferenzieren. Unter ‚Ausdifferenzierung‘ wird dabei verstanden, dass für das jeweilige Kompetenzniveau ganz genau beschrieben wird, was die Schülerinnen und Schüler können müssen, um auf diesem Niveau eingestuft zu werden.

Es müssen aber nicht nur testtheoretische, fachliche und fachdidaktische Kriterien beachtet werden, sondern die Kompetenzmodelle müssen auch bildungspolitischen Erwartungen und pädagogischen Erfordernissen entsprechen. Als Beispiel soll hier angeführt werden, wie diese Erwartungen und Erfordernisse für die Kompetenzmodelle der Bildungsstandard-Überprüfung formuliert werden (KMK, 2009):

- herausfordernde und zugleich angemessene Leistungserwartungen beschreiben, die der Leistungsstreuung innerhalb und zwischen den Bundesländern in angemessener Weise Rechnung tragen
- trotz der zu erwartenden unterschiedlich hohen Anteile von Schülerinnen und Schülern, die den Mindest-
  oder Regelanforderungen nicht entsprechen, für alle Länder ein ‚Leistungsminimum‘ beschreiben, das von allen Schülerinnen und Schülern mittelfristig erreicht wird
- vorhandene sowie auszubauende Leistungsressourcen verdeutlichen
- motivierende Leistungserwartungen formulieren, die Entwicklungsimpulse an den Schulen auslösen
- breite bildungspolitische Akzeptanz insbesondere bei den Lehrkräften erreichen
- in einer spannungsreichen Relation zu den internationalen Ergebnissen stehen.

Der weitere Bestandteil eines Kompetenzmodells, nämlich die Einteilung in Abstufungen, soll im nächsten Kapitel unter dem Begriff ‚Kompetenzniveau‘ betrachtet werden.

3.2.3.3 Kompetenzniveaus

Wie bereits beschrieben, ist es nicht nur wichtig zu untersuchen, aus welchen Teilen sich eine bestimmte Kompetenz in einem Kompetenzmodell zusammensetzt, sondern auch, wie diese Kompetenz abgestuft werden kann. Auf Grund dieser Kompetenzniveaus wird es ermöglicht, dass Testergebnisse kriterienorientiert interpretiert werden können (vgl. Kapitel 3.2.1.1).

Jedes Niveau ist durch „kognitive Prozesse und Handlungen von bestimmter Qualität" spezifiziert. Dabei sind dies immer Fähigkeiten, die die Schülerinnen und Schüler nur auf einem bestimmten Niveau, nicht aber auf einem darüber liegenden Niveau beherrschen (Klieme, 2007).

Beispiele solcher Stufenmodelle gibt es bei TIMSS und PISA, aber auch bei der Normierung der Bildungsstandards. Dabei wird davon ausgegangen, dass sich für jede Stufe bestimmte Aufgaben erstellen lassen, sodass Schülerinnen und Schüler auf Kompetenzniveau I die Aufgaben von Niveau I mit sehr großer Wahrscheinlichkeit lösen können, die Aufgaben auf Niveau II dagegen nur mit einer sehr geringen Wahrscheinlichkeit. Wie die einzelnen Stufen in Aufgaben umgesetzt werden, wird in Kapitel 3.3 beschrieben.

Im Folgenden sollen die Kompetenzniveaus von PISA im Bereich Mathematik[12] dargestellt werden. Hier wird deutlich, dass auf der unteren Kompetenzstufe Schülerinnen und Schüler zugeordnet sind, die „ein arithmetisches Wissen besitzen, das sie abrufen und anwenden können", auf der obersten Stufe dagegen „komplexe Modellierungen und mathematische Argumentationen" geleistet werden (Klieme, 2007).

Beispiel: Kompetenzniveaus in PISA: Mathematik

Kompetenzniveau I: Rechnen auf Grundschulniveau

Personen, die diesem Niveau zugeordnet werden, verfügen lediglich über arithmetisches und geometrisches Wissen auf Grundschulniveau. Sie können dieses Wissen abrufen und unmittelbar anwenden, wenn die Aufgabenstellung von vornherein eine bestimmte Standard-Mathematisierung nahe legt. Begriffliche Modellierungen sind nicht leistbar.

Kompetenzniveau II: Elementare Modellierungen

Auf diesem Niveau werden auch einfachste begriffliche Modellierungen vorgenommen, die in einem außermathematischen Kontext eingebettet sind. Personen auf diesem Kompetenzniveau können unter mehreren möglichen Lösungsansätzen den passenden finden, wenn durch Graphiken, Tabellen, Zeichnungen usw. eine Struktur vorgegeben ist, die das Modellieren erleichtert. Auch auf diesem Niveau sind allerdings nur die Wissensinhalte der Grundschulmathematik sicher verfügbar.

Kompetenzniveau III: Modellieren und begriffliches Verknüpfen auf dem Niveau der Sekundarstufe I

Mit diesem Niveau findet im Vergleich zu Niveau II in mehrfacher Hinsicht ein qualitativer Sprung statt. Schülerinnen und Schüler auf diesem Kompetenzniveau verfügen auch über einfache Wissensinhalte der Sekundarstufe I, also über den Standardstoff der Lehrpläne aller Schulformen. Sie können Konzepte aus unterschiedlichen mathematischen Bereichen verknüpfen und zur Lösung von Problemstellungen nutzen, wenn visuelle Darstellungen den Lösungsprozess unterstützen.

Kompetenzniveau IV: Umfangreiche Modellierungen auf der Basis anspruchsvoller Begriffe

Schülerinnen und Schüler auf diesem Kompetenzniveau bewältigen im technischen Bereich umfangreichere Verarbeitungsprozesse, können also eine Lösung über mehrere Zwischenergebnisse hinweg aufbauen. Auch offene Modellierungsaufgaben werden bewältigt, bei denen man unter vielfältigen Lösungswegen einen eigenen finden muss. Verstärkt können auch innermathematische begriffliche Zusammenhänge modelliert werden.

Kompetenzniveau V: Komplexe Modellierung und innermathematisches Argumentieren

Auf diesem letzten Niveau ist auch anspruchsvolles curriculares Wissen verfügbar. Die Schülerinnen und Schüler, die diesem Kompetenzniveau zugeordnet werden, können auch sehr offen formulierte Aufgaben bewältigen, bei denen ein Modell frei gewählt beziehungsweise selbst konstruiert werden muss. Begriffliche Modellierungsleistungen auf dieser höchsten Stufe umschließen häufig Begründungen und Beweise sowie das Reflektieren über den Modellierungsprozess selbst.

(nach: Klieme/Neubrand/Lüdtke, 2001)

12 Entsprechende Stufeneinteilungen existieren für alle in PISA getesteten Kompetenzen.

3.2.3.4 Standards

Auf Grund der Kompetenzmodelle und der Kompetenzniveaus ist es möglich, Standards jahrgangs- und fachbezogen festzulegen. Standards werden verbal formuliert und stützen sich auf die beschriebenen Modelle und Stufen. So können Mindest- oder Regelstandards festgehalten werden.

Definition: Mindeststandards

Mindeststandards beschreiben die Kompetenzen, die alle Schülerinnen und Schüler als Minimum erreichen sollen. Mindeststandards können erst dann formuliert werden, wenn die Standards bereits einen „Prozess der Erfahrung" durchlaufen haben. Wird dieser Prozess, in dem Aufgabenbeispiele validiert und Schwierigkeiten von Aufgaben getestet werden, nicht eingehalten, so kann es passieren, dass Schülerinnen und Schüler entweder über- oder unterfordert werden. Für die Mindeststandards sind Kompetenzniveaus eine unabdingbare Voraussetzung.

(Klieme, 2007; KMK, 2005)

Definition: Regelstandards

Regelstandards beschreiben die Fähigkeiten, die in der Regel von den Schülerinnen und Schülern erreicht werden sollen. Dabei werden die Fähigkeiten benannt, die von mindestens der Hälfte der Schülerinnen und Schüler erreicht werden sollen, also ein mittleres Anforderungsniveau. Die von der KMK formulierten Bildungsstandards sind beispielsweise Regelstandards. Sie basieren auf Einschätzungen von Experten aus Theorie und Praxis. Auch diese Standards müssen nach ihrer Einführung noch validiert werden.

(Klieme, 2007; KMK, 2005)

3.2.4 Weiterführende Literatur

Als Literatur zum ‚literacy'-Konzept empfiehlt sich Klieme/Baumert/Köller 2000. Auch der Aufsatz von Hanna Kiper (2003b) „literacy versus Curriculum?" gibt einen guten Einblick in das Konzept inklusive der Vor- und Nachteile. Zudem beschreiben die jeweiligen Berichte der einzelnen Studien (vgl. Kapitel 3.1.3.1 bis 3.1.3.4), ob und wie in ihnen das Konzept von ‚literacy' ein- und umgesetzt wird. Klieme (2007) beschreibt die Bereiche der Kompetenz im Rahmen der Standardentwicklung sehr ausführlich. Weitere Informationen über den Kompetenzbegriff in den jeweiligen Studien und die Einteilung in Kompetenzniveaus kann der Literatur zu den Studien entnommen werden (vgl. Kapitel 3.1).

3.2.5 Verständnis-Aufgaben und Diskussionspunkte

1. *Wer Leistungsvergleiche durchführen will, braucht Maßstäbe. Welcher Maßstab eignet sich für welche Problemstellung? Worin bestehen Vor- und Nachteile?*
2. *Konstruieren Sie einige Aufgaben in einem gewählten Fachbereich auf unterschiedlichen Kompetenzniveaus und dokumentieren Sie den Prozess.*
3. *‚Literacy'-Orientierung und curriculare Orientierung sind unterschiedliche Konzepte bei der Konstruktion eines Untersuchungsdesigns. Welche möglichen Auswirkungen hat die Auswahl auf die Schul- und Unterrichtsentwicklung?*
4. *Es gibt sowohl in der Wissenschaft als auch in der Politik eine Diskussion, die statt der/ergänzend zu den Regelstandards die Definition von Mindeststandards fordert. Reflektieren Sie diese Position und ihre Konsequenzen für das Bildungswesen.*

3.3 Testkonstruktion

Kapitel 3.3 beschäftigt sich mit der Testkonstruktion. Zunächst wird der theoretische Hintergrund beleuchtet, es folgen Betrachtungen der Testgütekriterien, der Test-Verfahrensweisen, der Test-Aufgabenauswahl und der Zielpopulation. Das Kapitel endet mit einer Übersicht über die bisher durchgeführten und geplanten Erhebungen im Rahmen des Systemmonitorings.

In diesem Kapitel werden folgende Fragen beantwortet:

- *Welchen Anforderungen müssen die Aufgaben der Studien genügen? Und warum?*
- *Wie verläuft die Testdurchführung?*
- *Wie erfolgt die Auswertung beziehungsweise Analyse der Ergebnisse?*

- *Wie werden Aufgaben für die Studie ausgewählt und wie werden sie in den Kompetenzmodellen verortet?*
- *Wie werden die Schülerinnen und Schüler ausgesucht, die an der Testdurchführung teilnehmen?*
- *Welche Studien wurden bereits durchgeführt und welche sind in Planung?*

Auch wenn die Zugänge zu diesen Fragestellungen eher theoretischer Natur sind, dienen sie im Besonderen dem Verständnis der Systemmonitoring-Studien. Von herausragender Bedeutung ist die Betrachtung der „professionellen Qualitätsmaßstäbe", um der Gefahr von Fehl- und Überinterpretationen zu begegnen (Klieme u.a., 2007).

Die in diesem Kapitel gegebenen Informationen sind wichtig, da so das Prinzip verstanden werden kann, das hinter den Systemmonitoring-Studien steht. Zudem ist eine Interpretation der Ergebnisse einfacher, wenn die Hintergründe bekannt sind, unter denen die erzielten Punktwerte der Schülerinnen und Schüler entstanden sind.

Im gesamten Kapitel sollen dabei ausschließlich die international durchgeführten Studien betrachtet werden.

3.3.1 Testtheoretischer Hintergrund

Leistung = beobachtbares Verhalten

Die wichtigste Voraussetzung für das Testen schulischer Leistungen ist, dass die Leistung sich durch direkt beobachtbares Verhalten festhalten lässt. Dies können sprachliche Äußerungen oder Aufgabenbearbeitungen sein.

Die Untersuchung muss so gestaltet werden, dass die unterschiedlichen Leistungspotenziale der Schülerinnen und Schüler dadurch erkennbar werden, dass sie sich ihrem Leistungsniveau entsprechend unterschiedlich verhalten. Diesen Leistungsniveaus werden bestimmte Skalenwerte zugeordnet. Für die Zwecke des Systemmonitorings wird versucht, ein Ordinal- oder Intervallskalenniveau (vgl. Erklärung unten) zu verwenden, da die Unterschiede zwischen den einzelnen Messwerten so möglichst aussagekräftig sind (Heller/Hany, 2001).

Vertiefung: Ordinal- und Intervallskala (optional)

Definition: Ordinalskala

„Die relative Größe der den Objekten zugeordneten Zahlen reflektiert die Ausprägung des Merkmals, die die Objekte besitzen. Gleiche Differenzen zwischen den Zahlen implizieren nicht gleiche Differenzen in den Merkmalsausprägungen."

Beispiele: Rangplätze im Sport, Schulnoten, Rangbildung hinsichtlich eines psychologischen Merkmals, Testrohwerte, Härte von Mineralien

(Diehl/Kohr, 2004)

Definition: Intervallskala

„Es existiert eine Maßeinheit, mit der die Objekte nicht nur zugeordnet werden können – man kann ihnen auch Zahlen so zuweisen, dass gleiche Differenzen zwischen den den Objekten zugeordneten Zahlen gleiche Differenzen in der Ausprägung des gemessenen Merkmals reflektieren. Der Nullpunkt ist willkürlich gesetzt und bedeutet nicht ‚Nichtvorhandensein' des Merkmals."

Beispiele: Kalenderzeit, Celsius und Fahrenheit, Temperaturskalen, Intelligenztestwerte

(Diehl/Kohr, 2004)

Probabilistische Testtheorie / Rasch-Modell

Es kann ein Modell der probabilistischen Testtheorie, das sogenannte Rasch-Modell (vgl. Studienbrief-Teil Vergleichsarbeiten), angewendet werden, um Fähigkeiten der Schülerinnen und Schüler und Schwierigkeiten der Items auf der gleichen Skala abzubilden. Übersteigt die Fähigkeit eines Schülers die Schwierigkeit eines Items, kann man davon ausgehen, dass dieser Schüler dieses Item mit größerer Wahrscheinlichkeit richtig beantwortet, als dass er es falsch beantwortet.

Proficiency Scaling

Im Bereich der Kompetenzdiagnostik wird die Erfassung von Kompetenzen mit Hilfe der probabilistischen Testtheorie auch als ‚Proficiency Scaling' bezeichnet. Die Testaufgaben werden dazu von Expertinnen und Experten des jeweiligen Inhaltsbereichsbezüglich ihrer Anforderungsmerkmale analysiert

(vgl. Studienbrief-Teil Vergleichsarbeiten). Aufgaben-Cluster mit ähnlichen Anforderungsmerkmalen und ähnlicher Schwierigkeit werden im Anschluss als „inhaltlich definierte Stufen der Kompetenz“ (Bos u.a., 2003) beschrieben. Aufgabenbeispiele ermöglichen einen plastischen Eindruck der Anforderungen, denen Schülerinnen und Schüler auf dem jeweiligen Niveau gerecht werden (vgl. für eine detaillierte Beschreibung Studienbrief-Teil Vergleichsarbeiten).

3.3.2 Testgütekriterien und Testanforderungen

Für Informationen zur klassischen Testtheorie sowie eine ausführliche Beschreibung der drei grundlegenden Testgütekriterien Objektivität, Reliabilität und Validität sei auf die Ausführungen in Studienbrief-Teil ‚Individuelle Diagnostik‘ verwiesen. In diesem Kapitel sollen die Testgütekriterien nur in aller Kürze und jeweils mit einem Bezug zu den Studien des Systemmonitorings dargestellt werden. Dabei muss beachtet werden, dass die Bewertung eines Testes anhand der Testgütekriterien immer verbunden sein muss mit dem Blick auf die Schlussfolgerungen, die daraus gezogen werden sollen sowie deren Art und deren Nutzung (Klieme u.a., 2007).

Validität

Besonders für die internationalen Studien des Systemmonitorings ist es bedeutsam, dass Untersuchungsgegenstände ausgewählt werden, die Auskunft über die Leistungsfähigkeit der jeweiligen Bildungssysteme geben. Außerdem müssen Voraussetzungen, Durchführung und Auswertung der Tests für alle Teilnehmerstaaten gleich sein. Die Betrachtung der Validität soll sicherstellen, dass genau das gemessen wird, was auch gemessen werden soll (Heller/Hany, 2001). Dazu werden die Studien im Rahmen des Systemmonitorings sorgfältig konzipiert und die Aufgaben gewissenhaft ausgesucht. Zudem durchlaufen die Testaufgaben eine Erprobungsphase.

Objektivität

Um dem Kriterium der Objektivität gerecht zu werden, erhalten bei Studien des Bildungsmonitorings alle Testleiter von den Organisationen und Instituten, die die Studie konzipiert haben, genaue Anweisungen zur Durchführung der Studie. Darin ist genau beschrieben, wie die Aufgaben verteilt werden müssen, welche Anweisungen gegeben werden dürfen und welche Hilfestellungen und Hilfsmittel erlaubt sind. Die Einhaltung der Anweisungen wird von unangemeldeten Beobachtern überwacht.

Testauswertung

Auch die Testauswertung muss konsistent durchgeführt werden, sodass gleiche Antworten zu einer gleichen Interpretation und Punktevergabe führen. Bei den großen Studien des Systemmonitorings wird die Testauswertung nicht von den Lehrerinnen und Lehrern, sondern von den Organisationen, die den Test entwickelt haben, durchgeführt. Hier werden Personen extra geschult um Ergebnisse zu erreichen, die dem Testgütekriterium der Objektivität entsprechen (Heller/Hany, 2001).

Reliabilität

Die Reliabilität fordert, dass die Messungen präzise und unabhängig von Ort oder Zeit der Testdurchführung sind. Theoretisch müssen sich die Tests wiederholen lassen und das gleiche Ergebnis zeigen. Auch aus diesem Grund müssen die Anweisungen an die durchführenden Personen möglichst genau sein. Da es aber unmöglich ist, Messfehler zu vermeiden, ist es möglich und sinnvoll abzuschätzen, wie groß der Anteil der Messfehler sein kann (vgl. Studienbrief Teil 1).

3.3.3 Test-Verfahrensweisen

Zeitfenster

Die meisten Systemmonitoringstudien finden an zwei Tagen hintereinander statt, da die Ergänzungsstudien meist einen Tag nach den internationalen Erhebungen stattfinden. Für die internationale Durchführung wird ein Zeitfenster festgelegt, in dem die Durchführung stattfinden muss. Dieses Zeitfenster betrug beispielsweise bei PISA 2006 einen Monat (PISA-Konsortium, 2007). Speziell geschulte Testleiterinnen und Testleiter lesen den Schülerinnen und Schülern die Anweisungen vor und beaufsichtigen die Durchführung des Tests. Alle Aktivitäten der Testleiterinnen und Testleiter sind vorgegeben und müssen genau befolgt werden. Dies reicht von der Begrüßung über das Verteilen der Testhefte bis zum Erklären und Einsammeln der Testhefte (Bos u.a., 2003).

Teil-Tests / rotierende Testhefte

Das Rasch-Modell (vgl. Kapitel 3.3.1) ermöglicht es, einen an Items umfangreichen Test in zahlreiche Untertests aufzuteilen, welche die gleiche Kompetenz messen und von den Schülerinnen und Schülern zeitökonomischer bearbeitet werden können. Die Untertests werden im Rahmen der Testauswertung wieder zusammengefügt; die jeweiligen Kompetenz-(Fähigkeits-)Schätzungen auf Basis unterschiedlicher Testhefte ermöglichen es dennoch, alle Schülerinnen und Schüler auf der gleichen Skala abzubilden. Dieses Verfahren mit ‚rotierenden Testheften‘ entlastet die Schülerinnen und Schüler, da sie nur eine begrenzte Anzahl von Aufgaben bearbeiten müssen, ohne dass Nachteile für die Studie entstehen, da diese weiterhin eine hohe Repräsentativität erhalten kann. Zudem wird das Abschreiben der Schülerinnen und Schüler beinahe unmöglich gemacht, da sie unterschiedliche Aufgaben bearbeiten (Arnold, 2001).

Multi-Matrix-Designs

Dieses Vorgehen nennt sich ‚Multi-Matrix-Design‘ oder ‚Multi(ple)-Matrix Sampling‘ und hat sich in der empirischen Bildungsforschung als Forschungsdesign etabliert. Die Teil-Tests oder Testhefte (bei IGLU

2001 gab es beispielsweise 12 verschiedene) ermöglichen es, dass die Schülerinnen und Schüler unterschiedliche Aufgaben bearbeiten (Bos u.a., 2003). Durch dieses Vorgehen ist es möglich, dass die Schülerinnen und Schüler, wie beispielsweise in PISA 2006, nur 120 Minuten an den Aufgaben arbeiteten, sie jedoch als Gemeinschaft Aufgaben bearbeiteten, die 390 Minuten in Anspruch genommen hätten. Die Kompetenzen der Schülerinnen und Schüler konnten auf diese Art und Weise wesentlich umfangreicher untersucht werden. Allerdings verlangt ein solches Vorgehen auch eine komplexe statistische Auswertung (PISA-Konsortium, 2007).

Vorteil

Komplexe statistische Auswertung

Vertiefung: Beispiel für Umgang mit rotierenden Testheften/Multi-Matrix-Design (optional)

Als Beispiel soll die Vorgehensweise bei PISA 2006 beschrieben werden. Hier erhielten die Schülerinnen und Schüler 13 unterschiedliche Testhefte. Dabei waren in jedem Testheft vier Aufgabengruppen vorhanden, die auch in drei weiteren Testheften vorkamen, sodass sich die Schwierigkeitsgrade der Aufgaben in den Aufgabengruppen unabhängig von den Testheften miteinander in Beziehung setzen ließen. Daraus folgt, dass in PISA 2006 jeder nur vier Dreizehntel der Aufgaben bearbeitete. Da sich die durchschnittlichen Schwierigkeiten der einzelnen Testhefte zwar ähnelten, aber nicht übereinstimmen, kann nicht die Anzahl der richtigen Lösungen als Maß für die Kompetenz der Schülerinnen und Schüler herangezogen werden. Daher wird auf Testmodelle der ‚Item-Response-Theorie' (IRT; vgl. Studienbrief-Teil Vergleichsarbeiten) zurückgegriffen, um die einzelnen Kompetenzwerte zu schätzen. Dabei geht man wie folgt vor:

Antwort-Modell/ Item-Response-Theorie (IRT)

> „Man postuliert eine nicht direkt beobachtbare (latente) kontinuierliche Personenvariable, die das Antwortverhalten der Schülerinnen und Schüler erklärt. Zum Beispiel kann man annehmen, dass die naturwissenschaftliche Kompetenz eine latente Variable darstellt, mit der die Antworten auf die Aufgaben des PISA-Naturwissenschaftstests erklärt werden können" (PISA-Konsortium, 2007).

Formalisiert werden kann dies durch ein Modell, in PISA ist dies das Rasch-Modell., Der größte Vorteil eines IRT-Modells im Vergleich zur klassischen Testtheorie (vgl. Studienbrief-Teil Individuelle Diagnostik) liegt darin, dass die Leistungen der Schülerinnen und Schüler auf einer gemeinsamen Skala abgebildet werden können, obwohl sie unterschiedliche Aufgaben bearbeitet haben (PISA-Konsortium, 2002). Weitere Eigenschaften eines solchen Skalierungsverfahrens, besonders für Studien des Systemmonitorings, sind zusammengefasst:

Vorteil des Rasch-Modells

- Es ermöglicht, dass die inhaltliche Passung von Testaufgaben für die theoretisch bestimmte Fähigkeitsdimension empirisch geprüft werden kann und ungeeignete Aufgaben aussortiert werden können.
- Es ermöglicht, die Fähigkeit einer Person zuverlässig zu schätzen, auch wenn nur eine Untermenge der insgesamt verfügbaren Aufgaben bearbeitet wurde (s. oben).
- Es ermöglicht, Teilnehmergruppen mit unterschiedlicher Fähigkeit maßgeschneiderte Testversionen zu geben, da die Schwierigkeiten der bearbeiteten Aufgaben bei der Schätzung der Fähigkeit von Personen berücksichtigt werden.
- Es ermöglicht, die Schwierigkeit von Testaufgaben und die Fähigkeit von Personen auf demselben Maßstab abzubilden.

Allerdings kann das (dichotome) Rasch-Modell nur für Aufgaben verwendet werden, bei denen es nur richtige oder falsche Antworten gibt. Gibt es Aufgaben, bei denen die Schülerinnen und Schüler mehr als einen Punkt erlangen können, etwa weil für eine Aufgabe auch Teillösungen bewertet werden, muss ein anderes Modell gewählt werden. In diesem Fall kann das ‚Partial-Credit-Modell' (ordinales Rasch-Modell) für ordinale Items verwendet werden (PISA-Konsortium, 2007). Nähere Informationen dazu können der entsprechenden Literatur entnommen werden.

Die Untersuchungen bei IGLU dauerten an beiden Tagen etwa zwei Stunden (Bos u.a., 2003). Dies ist auch für die anderen Studien ein ungefährer Richtwert, es sollte darauf geachtet werden, dass die Bearbeitungsdauer altersangemessen ist. Wichtig ist es, dass, gerade bei jüngeren Schülerinnen und Schülern, abwechslungsreiche Aufgaben angeboten werden und die Möglichkeit besteht, eine Aufgabe zu überspringen und an einer anderen Stelle weiterzumachen (Bos u.a., 2003).

Dauer

Zum Test gehören nicht nur die jeweiligen Aufgaben, sondern auch Fragebögen, die zusätzliche und ergänzende Informationen liefern können, um Leistungsunterschiede zu erklären. Diese werden von den

Fragebögen

Schülerinnen und Schülern im Rahmen der Testdurchführung ausgefüllt, die Eltern, Lehrerinnen und Lehrer und die Schulleitung füllen die Bögen ebenfalls während der zwei Testtage aus. Alle Unterlagen werden gemeinsam zur durchführenden Organisation zurückgeschickt (Bos u.a., 2003).

3.3.4 Testaufgabenauswahl

Inhalte

Die Tests unterscheiden sich darin, welche Kompetenzen die Studie untersucht. Vor Beginn der Studie werden die zu untersuchenden Inhalte festgelegt, dabei kann ein bestimmtes Thema besonders ausführlich oder nur sehr knapp in die Planung eingehen. Bei Studien, die der ‚literacy'-Orientierung folgen, orientieren sich die Inhalte nicht an Lehrplänen. Bei PISA weichen beispielsweise die Aufgaben davon ab, was die Schülerinnen und Schüler aus dem Unterricht kennen, und zwar sowohl in Bezug auf die Aufgabenstellung und die damit verbundenen Anforderungen als auch auf die Themen und Inhaltsbereiche. So ist es nicht immer möglich, die Aufgaben aus den Studien direkt einem Schulfach zuzuordnen, da beispielsweise die Lesekompetenz als fächerübergreifende Kompetenz auch in den anderen Fächern involviert ist (Deutsches PISA-Konsortium, 2003). Die Studien versuchen, die Aufgaben möglichst nah am Alltag der Schülerinnen und Schüler anzusiedeln und trotzdem die fachlichen Kenntnisse zu überprüfen (Baumert/Bos/Watermann, 1999). Eine Veröffentlichung der Aufgaben nach Abschluss der Studien kann nur in sehr begrenztem Umfang stattfinden, da die Aufgaben aus methodischen Gründen für die nachfolgenden Erhebungswellen und einen längsschnittlichen Vergleich benötigt werden (van Ackeren, 2006).

Aufgabentypen

Die Aufgabentypen unterscheiden sich danach, welche Kompetenzen der Schülerinnen und Schüler getestet werden sollen. Auf Grund einer größeren Auswertungsobjektivität werden meist geschlossene Aufgaben ausgewählt, die nur eine richtige Antwort haben. Die bekanntesten und auch meist verwendeten Aufgaben sind dabei ‚Multiple-Choice-Aufgaben' (Mehrfachwahlaufgaben). Aber auch offene Aufgabenformate können gestellt werden, wie beispielsweise in IGLU (Bos u.a., 2003). Zu diesen offenen Aufgabenformaten gehören Aufgaben, die Kurzantworten verlangen (ein einziges Wort, mehrere Wörter, Zeichnungen). Meist werden die einzelnen Aufgaben in Gruppen zusammengefasst, wobei sich jeweils eine Gruppe der Aufgaben mit der Beschreibung einer bestimmten alltagsnahen Situation beschäftigt (PISA-Konsortium, 2002).

Rückmeldung

Aufgabenentwicklung

Im Rahmen der Aufgabenentwicklung werden zunächst Aufgaben gesammelt, die für den Zweck der Studie theoretisch einsetzbar wären. Hierbei kann einerseits auf Aufgaben zurückgegriffen werden, die bereits in anderen Studien verwendet wurden, andererseits werden internationale Experten dazu aufgerufen, geeignete Aufgaben einzusenden. Beispielsweise für TIMSS wurde eine Aufgaben-Datenbank für die Aufgaben angelegt. Anschließend werden die Aufgaben von Experten begutachtet und die ausgewählten Aufgaben ausprobiert (Baumert/Lehmann, 1997). Zu beachten ist allerdings, dass TIMSS eine curriculare Orientierung verfolgt und daher die curriculare Validität durch diese Überprüfung sichergestellt werden sollte. Bei der Überprüfung der curricularen Validität der TIMSS-Aufgaben wurde auf die folgenden Kriterien geachtet (Baumert/Bos/Lehmann, 2000):

- Lehrplanvalidität des in einer Aufgabe repräsentierten Stoffes
- Vertrautheit mit der spezifischen Einbettung und Präsentation des Stoffes
- fachliche Qualität der Aufgabe unabhängig von ihrer curricularen Gültigkeit
- vermutete Lösungswahrscheinlichkeit der Aufgabe

Umgang mit unbekannten Aspekten

Eine solche Überprüfung der Angemessenheit der Aufgaben muss für Studien mit ‚literacy'-Orientierung entsprechend angepasst werden. Wichtig ist, dass keine unbekannten Aspekte in den Aufgaben enthalten sind. Dies ist vor allem von großer Bedeutung, wenn die Schülerinnen und Schüler noch recht jung sind, da sie ansonsten überfordert sind und die Testdurchführung abbrechen könnten[13] (Bos u.a., 2008).

Erprobung der Aufgaben

In Pre-Tests[14] vor Beginn der eigentlichen Untersuchung werden die Aufgaben dann auf ihre Tauglichkeit untersucht.

Vertiefung: Definition von Kompetenzniveaus (optional)

Definition von Kompetenzniveaus

Durch die Nutzung der probabilistischen Testtheorie und der IRT-Skalierung ist es möglich, unterschiedliche Kompetenzniveaus zu definieren. In TIMSS III (1995) wurde beispielsweise so verfahren, dass zu-

13 Eine solche Beurteilung, ob ein Aspekt bekannt oder unbekannt ist, kann von Experten durchgeführt werden, die beispielsweise mit Hilfe von Schulbüchern (bei curricularer Orientierung) eine solche Einschätzung treffen. Dabei kann es sich um unbekannte graphische Darstellungsformen, mathematische Begriffe, Sachverhalte, Verfahren und Bezeichnungen handeln (Bos u.a., 2008).

14 In Pre-Tests werden die geplanten Erhebungsinstrumente erprobt, die Ergebnisse fließen nicht mit in die Studie ein. Ziel ist es, dass für die eigentliche Untersuchung Fehlerquellen im Vorfeld erkannt und ausgeräumt werden können.

nächst Fixpunkte auf der Skala ausgewählt wurden. Um anschließend einen dieser Punkte als ‚Kompetenzniveau' zu definieren, wurden die Aufgaben begutachtet, die die Personen mit diesem Punktwert erreicht haben, mit hinreichender Sicherheit (65 Prozent) lösen konnten, nicht aber von Personen mit Punktwerten auf dem niedrigeren Niveau. Die für diese Aufgabe charakteristischen Anforderungen beschreiben dann das Kompetenzniveau (Baumert/Bos/Watermann, 1999).

Zudem ist es durch die Skalierung möglich, jeder Aufgabe einen Schwierigkeitswert zuzuordnen und damit einem Kompetenzniveau. Als Beispiel soll hier die Verteilung von drei Aufgaben auf die Kompetenzniveaus aus IGLU-E 2001 dienen.

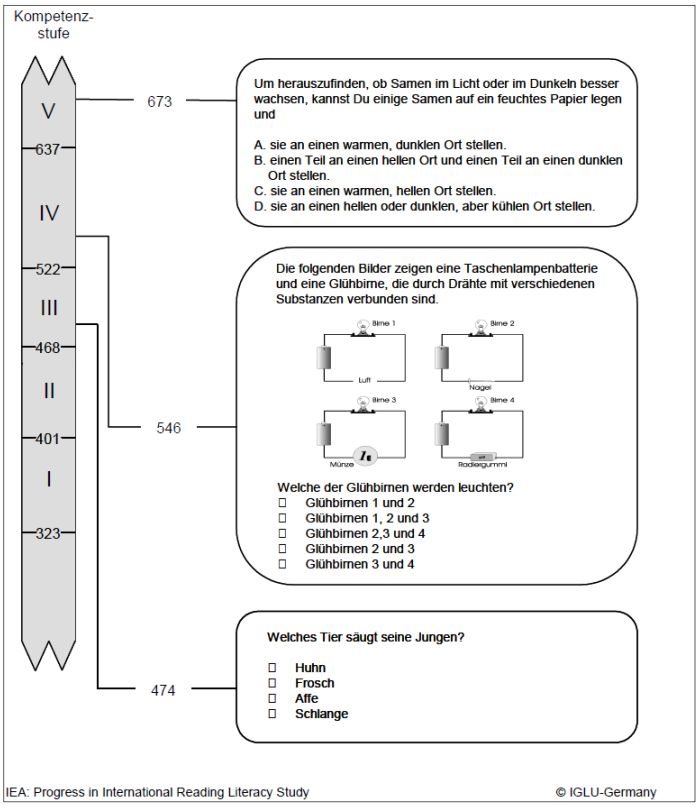


Abbildung 4: Verortung der Aufgaben nach Kompetenzniveaus bei IGLU

Das Beispiel macht deutlich, wo auf der Skala die Aufgaben ihrer Schwierigkeit nach anzuordnen sind.

Verankerung des Maßstabs

Erst durch die Verknüpfung eines Niveaus mit einer Beschreibung der zu vollziehenden Aufgaben und Operationen kann der Maßstab des Niveaus verankert werden und eine Aussagekraft erhalten (Baumert/Lehmann, 1997). Eine solche Beschreibung wurde bereits in Kapitel 3.2.3.3 gegeben. Es handelt sich hier um eine kriteriale Bezugsnorm (vgl. Kapitel 3.2.1.1).

Alle Schülerinnen und Schüler werden ihrem Ergebnis nach in eines der Kompetenzniveaus eingeteilt. Dadurch ist es möglich, den numerischen Wert des Ergebnisses inhaltlich zu verankern und ihm damit eine Aussagekraft zu verleihen (Baumert/Bos/Watermann, 1999).

Kompetenzniveau	Beschreibung	Skalenbereich der Fähigkeit
I	Gesuchte Wörter in einem Text erkennen	375-450
II	Angegebene Sachverhalte aus einer Textpassage erschließen	451-525
III	Implizit im Text enthaltene Sachverhalte aufgrund des Kontextes erschließen	526-600
IV	Mehrere Textpassagen sinnvoll miteinander in Beziehung setzen	> 600

Tabelle 8: Beschreibung der Kompetenzniveaus beim Leseverständnis und ihre Skalenwerte in IGLU (nach Bos u.a., 2003)

Kompetenzniveaus

Häufig werden vier Kompetenzniveaus definiert. Das folgende Beispiel stammt aus IGLU und zeigt die vier Kompetenzniveaus, die Beschreibung der auf dieser Stufe erreichten Kompetenzen und den Skalenbereich, innerhalb dessen die Schülerinnen und Schüler liegen müssen, um einem bestimmten Kompetenzniveau zugeteilt zu werden. Entsprechende Beschreibungen werden in allen Studien des Systemmonitorings erstellt.

Wichtig ist dabei, dass die Ergebnisse der Skalierung in den Studien durch eine Transformation so standardisiert werden, dass alle Studien einen Mittelwert von 500 Punkten und eine Standardabweichung von 100 Punkten aufweisen.[15]

Definition: Standardabweichung

Die Standardabweichung kann als „Maß der durchschnittlichen Abweichung vom Mittelwert" definiert werden. Aus ihr kann abgelesen werden, wie weit sich die Werte um den Mittelwert herum verteilen. Ein hoher Wert sagt aus, dass die Werte weit um den Mittelwert herum verteilt sind, ein kleiner Wert, dass die Werte nah am Mittelwert liegen. Zudem sagt die Standardabweichung auch etwas über die Reichweite der Wertverteilung aus (Albert/Koster, 2002).

Problem: Übersetzung der Aufgaben

Bei internationalen Untersuchungen muss darauf geachtet werden, dass alle Aufgaben in alle Sprachen gleich übersetzt werden. Darum werden beispielsweise bei IGLU die Aufgaben und auch die Fragebögen vom Englischen in die jeweilige Unterrichtssprache und wieder zurück übersetzt. Es wird kontrolliert, ob die Aufgaben dann noch immer mit der Ausgangsaufgabe übereinstimmen. Zudem werden die Aufgaben von Experten der Bundesländer auf ihre curriculare (inhaltliche) Validität begutachtet. Außerdem wird begutachtet, ob die Aufgaben angemessen in den Bereichen Sprache, Inhalt, Bearbeitungszeit und Vertrautheit mit der Textsorte (bei IGLU: Lesekompetenz) sind (Bos u.a., 2003).

Fragebögen

Innerhalb der Untersuchung werden aber nicht nur fachliche Inhalte untersucht, sondern auch weitere Faktoren, die die Leistungen der Schülerinnen und Schüler näher erklären sollen (Determinanten von Schulleistungen, vgl. Helmke & Schrader (1998), Studienbrief-Teil Vergleichsarbeiten). Vor allem internationale Unterschiede können so genauer betrachtet werden. Dazu werden Fragebögen konzipiert, die einerseits von den Schülerinnen und Schülern und ihren Eltern, andererseits aber auch von den Lehrerinnen und Lehrern oder den Schulleitungen ausfüllen werden. Welche Bereiche in diesen Fragebögen abgefragt werden, soll hier am Beispiel von IGLU dargestellt werden (Bos u.a., 2003):

- Familiärer Kontext
  - leseunterstützende Aktivitäten der Eltern
  - Sprache in der Familie
  - familiäre Ressourcen
  - Beziehungen zwischen Elternhaus und Schule
  - außerschulische Leseaktivitäten der Kinder
- Schulischer Kontext
  - Schulstandort und schulische Ressourcen
  - Lehrerausbildung
  - Lernumgebung und Unterrichtsorganisation in der Klasse
  - Unterrichtsmethoden und -maßnahmen
  - Unterrichtsmaterial und -medien
- Nationaler und kommunaler Kontext
  - demographische Besonderheiten
  - Führung und Organisation des Bildungssystems
  - Lehrplanbesonderheiten und -schwerpunkte

Sowohl die Informationen aus den Fragebögen als auch aus den Testaufgaben werden streng vertraulich behandelt, alle Aspekte des Datenschutzes werden berücksichtigt. Die so erlangten Informationen dürfen nur zu wissenschaftlichen Zwecken verwendet werden (van Ackeren, 2006).

3.3.5 Bezugsgruppen

Damit verschiedene Gruppen von Schülerinnen und Schülern international verglichen werden können, müssen nicht nur die Inhalte, sondern auch die Stichproben der Populationen vergleichbar sein.

15 Bei einer wiederholten Durchführung der Studie mit denselben Test-Items wird dann allerdings der tatsächlich errechnete Wert als Standard verwendet.

Ausmaß der Selektivität beachten

In den großangelegten internationalen Studien des Systemmonitorings ist es nicht möglich, die Grundgesamtheit der Schülerinnen und Schüler zu testen. Um Vergleiche zwischen Staaten oder Bundesländern ziehen zu können, muss das „Ausmaß der Selektivität der Stichproben, die gezogen und deren Kennwerte in den Vergleich genommen werden, abschätzbar" gemacht werden. Studien der Sekundarstufen müssen beispielsweise beachten, dass in einigen Staaten verschiedene Schulformen existieren. Im Vergleich zu Staaten mit integrativen Systemen muss bei solchen mit einem gegliederten Schulsystem „diese Struktur durch eine so genannte Stratifizierung (d.h. schulformbezogene Aufteilung der Stichprobe und anschließende Gewichtung der Kennwerte) sehr genau nachgebildet werden" (Arnold, 2001). Ebenfalls ist es problematisch, eine bestimmte Klassenstufe als Bestandteil der Population zu beschreiben, da das Alter der Schülerinnen und Schüler in einer bestimmten Klassenstufe international variiert. Daher wird vor Beginn der Untersuchung definiert, wer genau getestet werden soll. Eine Möglichkeit ist es, das Alter der teilnehmenden Schülerinnen und Schüler zu definieren (Beispiel: 15-Jährige, PISA) oder aber eine Klassenstufe, die dann näher beschrieben werden muss (Beispiel: Die Klassenstufe, in der die meisten Neunjährigen sind, IGLU).

Umgang mit Sonderschulen

Ein weiteres Problem entsteht, wenn die zu untersuchenden Schulsysteme unterschiedliche Quoten an Sonderschulen aufweisen und diese nicht mit in die Untersuchung einbezogen werden (Arnold, 2001). Aus diesem Grund werden in den internationalen Studien Quoten festgelegt, nach denen bestimmte Schülergruppen von der Untersuchung ausgeschlossen werden dürfen. Bei IGLU betrug diese Quote beispielsweise fünf Prozent der Population, ansonsten konnte das Land nicht in der internationalen Berichterstattung analysiert werden[16] (Bos u.a., 2003). Ausgeschlossen werden können auch Schülerinnen und Schüler, die die Testsprache nicht als Erstsprache sprechen und erst weniger als ein Jahr in dieser Sprache unterrichtet werden (PISA-Konsortium, 2007).

Umgang mit anderen Erstsprachen

Größe der Stichprobe

Generell ist es wichtig, dass bei internationalen Untersuchungen die Stichprobe groß genug ist. Dies resultiert vor allem aus dem komplexen Stichproben-Design. Deshalb werden von den durchführenden Organisationen Standards für die Stichprobenziehung festgelegt. Zudem kann vorgegeben werden, wie viele Schulen in der Stichprobe enthalten sein sollten. Bei IGLU waren dies beispielsweise 150 Schulen pro Land.

Ziehung der Schülerinnen und Schüler

Des Weiteren wurden die Stichproben in IGLU durch ein zweistufiges Verfahren geleitet. Zunächst wurden bundeslandweise Schulen gezogen, anschließend Klassen innerhalb dieser Schulen. Die Schulen wurden mit einer Wahrscheinlichkeit proportional zur Größe der Schule gezogen. Dies ist ein Verfahren, das in den meisten Studien des Systemmonitorings angewendet wird, da die Schulanzahl von Jahr zu Jahr als konstant angesehen wird und die Schülerzahlen ebenfalls keinen großen Schwankungen unterliegen. Für alle Klassen einer Schule besteht dann die gleiche Wahrscheinlichkeit, für die Stichprobe gezogen zu werden.

Durch diese Methode repräsentiert ein an der Studie teilnehmender Schüler nicht nur sich selbst, sondern auch eine bestimmt Anzahl anderer Schülerinnen und Schüler. Die Anzahl dieser Personen lässt sich dadurch berechnen, wie groß seine eigene Wahrscheinlichkeit ist, in die Stichprobe zu gelangen. Diese wiederum setzt sich aus der Wahrscheinlichkeit zusammen, dass die Schule gezogen wird und der Wahrscheinlichkeit, dass die Klasse innerhalb der Schule gezogen wird (Bos u.a., 2003). Möglich ist es aber auch, dass bestimmte Schulen oder Personen in der Stichprobe überrepräsentiert werden, ein solches Vorgehen nennt sich ‚oversampling'. Dies war beispielsweise in TIMSS III (1995) der Fall. Da zusätzliche Untersuchungen nur für die gymnasiale Oberstufe durchgeführt werden sollten, wurden Gymnasien stärker in die Stichprobe einbezogen. Das gleiche galt für die fünf am stärksten vertretenen Ausbildungsberufe an den Berufsschulen (Baumert/Bos/Watermann, 1999). Diese Überrepräsentation muss dann in der normalen Auswertung beachtet und relativiert werden.

Tatsächliche Teilnahme am Test

Wichtig ist zudem, dass die ausgewählten Schülerinnen und Schüler auch tatsächlich an der Untersuchung teilnehmen. Daher wurde beispielsweise bei PISA 2006 eine Mindestbeteiligungsquote von 80 Prozent gefordert. Trotzdem ist nicht auszuschließen, dass eher die leistungsschwachen Schülerinnen und Schüler der Untersuchung fernbleiben. Daher wurden im Rahmen von PISA 2003 und 2006 für alle gezogenen Schülerinnen und Schüler die entsprechenden Schulnoten festgehalten. Eine Analyse der Daten ergab, dass die Noten der teilnehmenden und der nicht-teilnehmenden Schülerinnen und Schüler

16 Werden in einem Land zu viele Schülerinnen und Schüler ausgeschlossen, die durch ihre Einschränkungen tendenziell schlechtere Ergebnisse erreicht hätten, so werden die Ergebnisse überschätzt (Baumert/Lehmann, 1997).

sich nur gering unterschieden und somit ausgeschlossen werden konnte, dass eine Ergebnisverzerrung vorliegt (PISA-Konsortium, 2007).

Soll eine nationale Ergänzungsstudie durchgeführt werden, so muss die Stichprobe erhöht werden, damit die dort vorhandene Fragestellung zufallskritisch abgesichert werden kann. Damit die Ergebnisse auf nationaler Ebene interpretiert werden können, werden die Bundesländer entsprechend der proportionalen Anteile der realen Schülerzahlen gewichtet (Bos u.a., 2003).

Ergänzungsstudien

3.3.6 Übersicht durchgeführter und anstehender Erhebungen (optional)

Die folgenden Tabellen geben einen Überblick darüber, welche Studien bereits durchgeführt wurden und welche noch durchgeführt werden. In den Tabellen ist vermerkt, welche Bezugsgruppe und welche Inhalte in den Studien getestet wurden beziehungsweise werden.

Wann?	Wer?	Was?	Studie
1995	Sekundarstufe I und II	Mathematik und Naturwissenschaften	TIMSS
2000	Sekundarstufe I	Lesen, Mathematik, Naturwissenschaften	PISA
2001	Primarstufe	Lesen, Mathematik	IGLU
2003	Sekundarstufe I	Lesen, Mathematik, Naturwissenschaften	PISA
2006	Primarstufe	Lesen, Mathematik	IGLU
2006	Sekundarstufe I	Lesen, Mathematik, Naturwissenschaften	PISA
2007	Primarstufe	Mathematik, Naturwissenschaften	TIMSS
2009	Sekundarstufe I	Lesen, Mathematik, Naturwissenschaften	PISA

Tabelle 9: Bereits durchgeführte Erhebungen von Studien des Systemmonitorings

Wann?	Wer?	Was?	Studie
2011	Primarstufe	Deutsch	IGLU
2011	Primarstufe	Mathematik	TIMSS
2012	Sekundarstufe I	Lesen, Mathematik, Naturwissenschaften	PISA
2015	Sekundarstufe I	Lesen, Mathematik, Naturwissenschaften	PISA
2016	Primarstufe	Deutsch, Mathematik	IGLU
2018	Sekundarstufe I	Lesen, Mathematik, Naturwissenschaften	PISA

Tabelle 10: Geplante Erhebungen von Studien des Systemmonitorings

Die Tabellen zeigen den geplanten Rhythmus der jeweiligen Studien: Alle drei Jahre wird PISA durchgeführt, alle fünf Jahre IGLU, alle vier Jahre TIMSS.

Plöner Beschlüsse

Eine weitere wichtige Entscheidung hat die KMK 2006 in ihren „Plöner Beschlüssen“ festgehalten: Die regelmäßige Durchführung des Systemmonitorings soll mit den Bildungsstandards verbunden werden. Dazu werden die oben beschriebenen Studien eingesetzt, um zu überprüfen, wie hoch der Anteil der Schülerinnen und Schüler ist, die die Bildungsstandards erreichen (Köller, 2008). Für die Primarstufe wird in den Fächern Deutsch und Mathematik, für den Hauptschulabschluss die Fächer Deutsch, Mathematik und erste Fremdsprache und für den mittleren Schulabschluss die Fächer Deutsch, Mathematik, erste Fremdsprache, Biologie, Chemie und Physik, jeweils an die internationalen Studien angelehnt, getestet.

3.3.7 Weiterführende Literatur

Einen genauen Einblick in die Testkonstruktion der einzelnen Studien geben die Berichte der Studien (siehe Kapitel 3.1.3.1 bis 3.1.3.4). Nähere Informationen zum Umgang mit Multi-Matrix-Designs bietet dabei vor allem die PISA-Literatur.

3.3.8 Verständnis-Aufgaben und Diskussionspunkte

1. *Warum sollten Aufgaben für eine Studie auf ihre Praxistauglichkeit untersucht werden?*
2. *Wie unterscheiden sich Aufgaben für eine Klassenarbeit von Testaufgaben?*
3. *Recherchieren Sie die Bedeutung des Begriffes „kumulatives Wissen“, setzen Sie ihn in den Zusammenhang mit der Kompetenzniveauentwicklung und entwickeln Sie ein Beispiel.*

3.4 Anwendungsbereich

Kapitel 3.4 beschreibt den Anwendungsbereich von Systemmonitoring-Studien. Dazu werden die Ergebnisse der in Kapitel 3 betrachteten Studien zusammengefasst, zuvor werden die notwendigen theoretischen Grundlagen beschrieben. Zudem soll beschrieben werden, in wie weit sich durch die Ergebnisse Hinweise auf Veränderungsnotwendigkeiten im deutschen Bildungssystem erkennen lassen.

In diesem Kapitel werden folgende Fragen beantwortet:

- *Welche theoretischen Grundlagen werden für eine Interpretation der Ergebnisse benötigt?*
- *Welche Ergebnisse und Interpretationen haben die bisherigen Erhebungen von TIMSS, PISA, IGLU und DESI gezeigt?*
- *Welche Aussagekraft haben diese Ergebnisse?*
- *Wie können die Ergebnisse im Schulalltag verwendet werden?*

3.4.1 Theoretische Grundlagen

Skalierung

Wie bereits im vorangegangenen Kapitel beschrieben, werden die Ergebnisse in den Systemmonitoringstudien skaliert, sodass alle Studien einen Mittelwert von 500 Punkten und eine Standardabweichung von 100 Punkten aufweisen.[17]

Streuung

Erreicht ein Schüler einen Punktwert von 500, so entspricht dieser der mittleren Kompetenz aller Schülerinnen und Schüler in den Teilnehmerstaaten. Die Ergebnisse der Schülerinnen und Schüler sind so standardisiert, dass 68,2 Prozent und damit etwa zwei Drittel mit ihrem Punktwert zwischen den Punktwerten 400 und 600, also eine Standardabweichung über und eine Standardabweichung unter dem Mittelwert, liegen. Erhöht man dies auf jeweils zwei Standardabweichungen, so liegen zwischen einem Punktwert von 300 und 700 Punkten 95,4 Prozent der Schülerinnen und Schüler mit ihrem Punktwert in diesem Bereich. Daraus folgt, dass nur 2,3 Prozent der an der Studie teilnehmenden Schülerinnen und Schüler einen Punktwert von unter 300 oder über 700 erreichen.

Internationale Vergleiche

Mittelwert und Standardabweichung helfen ebenfalls dabei, internationale Vergleiche anzustellen. Liegt beispielsweise der Mittelwert eines Staats bei 600 Punkten, so kann konstatiert werden, dass die Schülerinnen und Schüler in diesem Staat Ergebnisse erreichen, die deutlich über denen des internationalen Durchschnitts liegen. Für einen Vergleich kann auch die Standardabweichung eines einzelnen Staats hilfreich sein. Weist ein Land eine Standardabweichung auf, die größer ist als 100 Punkte, so streuen die Ergebnisse der Schülerinnen und Schüler in diesem Land stärker um den Mittelwert. Wäre die Standardabweichung kleiner, so hätten die Schülerinnen und Schüler homogenere Ergebnisse erreicht.

Kompetenzniveaus

Außerdem sind die Kompetenzniveaus wichtig für die Interpretation der Ergebnisse. Sie können dafür genutzt werden, um zu überprüfen, in wie weit sich die Verteilung auf die einzelnen Stufen in den Staaten unterscheidet. Zudem kann so der Anteil der Schülerinnen und Schüler ausgemacht werden, die schlechte Chancen für ihren weiteren Lebensweg haben, da sie sich auf dem untersten Kompetenzniveau oder sogar noch darunter befinden. Dies gilt auch umgekehrt für die Schülerinnen und Schüler, die besonders gut abschneiden und besonders gute Chancen haben (PISA-Konsortium Deutschland, 2007). Eine Betrachtung der Kompetenzniveaus bietet somit einen höheren Informationsgehalt als reine Vergleiche der Mittelwerte (Deutsches PISA-Konsortium, 2001).

3.4.2 Wesentliche Ergebnisse (optional)

In diesem Kapitel sollen die Ergebnisse der Studien TIMSS, PISA, IGLU und DESI vorgestellt werden. Eine solche Betrachtung bisheriger Ergebnisse verdeutlicht, wie zukünftige Ergebnisse interpretiert werden können und wie groß das Spektrum möglicher Interpretationen ist. Dabei sollen nicht nur reine Punktwerte sondern auch mögliche Erklärungen für diese Werte beschrieben werden. Das Hauptaugenmerk soll auf den internationalen Vergleichen liegen und dabei auf den Ergebnissen der in Deutschland getesteten Schülerinnen und Schüler. Da aber auch die Ergebnisse der Ergänzungsstudien wichtige Aussagen über Veränderungsnotwendigkeiten innerhalb Deutschlands liefern, sollen diese jeweils kurz betrachtet werden. Interessante Aspekte aus der Erhebung von Hintergrundinformationen (wie etwa zum sozialen Hintergrund der Schülerinnen und Schüler) sollen ebenfalls betrachtet werden. Da der Studienbrief-Teil

17 Bei einer wiederholten Durchführung der Studie mit denselben Test-Items wird dann der tatsächlich errechnete Wert als Standard verwendet.

nicht allumfassend über die Ergebnisse berichten kann und dies auch nicht seinem Ziel entspricht, sollen nur allgemeine und besonders auffällige Aspekte beschrieben werden. Dabei soll deutlich werden, dass nicht die reine Rangfolge, wie in der Presse häufig verwendet, wichtig ist, sondern vielmehr die jeweiligen erreichten Werte wie Mittelwert, Standardabweichung und die Verteilung auf die Kompetenzniveaus. Weitere Informationen zu den einzelnen Studien können der jeweils angegebenen Literatur entnommen werden.

3.4.2.1 TIMSS (optional)

Generell kann für die Leistungen der deutschen Schülerinnen und Schüler in TIMSS 1995 festgehalten werden, dass sie sowohl in Mathematik als auch in den Naturwissenschaften „in einem breiten internationalen Mittelfeld" liegen. Dies lässt sich sowohl für die Sekundarstufe I (TIMSS II) als auch für die Sekundarstufe II (TIMSS III) erkennen:

TIMSS II

- TIMSS II: Deutsche Schülerinnen und Schüler erreichen einen Mittelwert von 509 Punkten, wobei der internationale Durchschnitt bei 513 Punkten liegt, das bedeutet, dass beispielsweise mathematische Routineverfahren aus der sechsten bis achten Klasse „einigermaßen sicher" durchgeführt werden können.
- TIMSS II: Leistungen in den Naturwissenschaften besser als in Mathematik, allerdings nicht so große Unterschiede wie in anderen Staaten
- Besorgniserregend: Etwa 20 Prozent der Schülerinnen und Schüler in der achten Klasse sind im Bereich der Naturwissenschaften noch auf dem Niveau von Grundschulkenntnissen.
- TIMSS III: Defizite auch in der gymnasialen Oberstufe

TIMSS III

- Für alle Schulformen: Auch bei den Schülerinnen und Schülern der Sekundarstufe II ist im naturwissenschaftlichen Bereich das unterste Kompetenzniveau überrepräsentiert.
- Defizite besonders bei der Verknüpfung der fachlichen Kenntnisse mit der Anwendung bei alltagsnahen Problemen. Dies lässt auf eine im Unterricht fehlende Anbindung an die Alltagswelt der Schüler schließen.

Zur Veranschaulichung der Ergebnisse wird eine Tabelle dargestellt, wie sie in den Berichten der Studien meist vorhanden ist. Dabei wird für die vier Kompetenzniveaus (hier in ihrer ausformulierten Form von oben nach unten, für den Bereich Mathematik in TIMSS III) angegeben, wie viel Prozent der Schülerinnen und Schüler diese erreichen.

	Deutschland	Frankreich	Niederlande	Norwegen	Schweiz
Alltagsbezogene Schlussfolgerungen	15,4	0,0	3,7	0,6	0,8
Anwendung von einfachen Routinen	36,6	34,3	21,5	35,0	29,3
Bildung von Modellen und Verknüpfung von Operationen	34,1	47,6	41,3	40,2	43,0
Mathematisches Argumentieren	13,9	18,1	33,4	24,2	26,9

Tabelle 11: Verteilung der Schülerinnen und Schüler aus Deutschland, Frankreich, Niederlande, Norwegen und der Schweiz auf die vier Kompetenzniveaus des Bereichs Mathematik

Hier wird deutlich, dass Deutschland im untersten Niveau überrepräsentiert (15,4 Prozent) und im obersten Niveau unterrepräsentiert ist (13,9 Prozent).

Im internationalen Vergleich von TIMSS II **wurde** die folgenden Aspekte deutlich:

Schüler älter

- Die deutschen Schülerinnen und Schüler erreichen das Niveau der internationalen Mittelgruppe (Deutschland und 11 weitere Staaten mit ähnlichen Ergebnissen) erst, wenn sie sechs bis zwölf Monate älter sind als die Schülerinnen und Schüler aus den anderen Staaten dieser Gruppe (durch spätere Einschulungszeitpunkte, Zurückstellungen und eine hohe Quote von Wiederholern), diese Entwicklung zeigte sich auch in TIMSS III.

- Am besten kann Deutschland in TIMSS II mit den angelsächsischen Staaten verglichen werden, da die Schülerinnen und Schüler der nord-, ost- und westeuropäischen sowie asiatischen Staaten einen Leistungsvorsprung von mehr als einem Jahr aufweisen können. Vergleichbare Länder
- Im Bereich der Mathematik bilden diese Staaten eine Leistungsgruppe, die Punktwerte erreichen konnte, die etwa eine halbe Standardabweichung über den deutschen Ergebnissen liegt, dies entspricht in etwa dem Leistungsfortschritt aus einem Schuljahr. Leistungsgruppe Mathematik
- Von den besten zu den schlechtesten Schülerinnen und Schülern besteht ein großer Unterschied von mehreren Schuljahren, sodass dieser Vorsprung praktisch nicht mehr eingeholt werden kann. Dies wird vor allem dadurch verstärkt, dass sich die Unterschiede zwischen den guten asiatischen Staaten vom Beginn der Sekundarstufe I zum Ende der Sekundarstufe I beispielsweise im Bereich Mathematik von knapp einer Standardabweichung auf über eine Standardabweichung nochmals vergrößern (auch durch TIMSS III wird erkennbar, dass der Abstand größer wird) Entwicklung über ein Schuljahr

In Bezug auf die Streuung der Punktwerte lassen sich die folgenden Aspekte zusammenfassen:

- Streuung der Punktwerte bei deutschen Schülerinnen und Schülern besonders groß, große Heterogenität in beiden getesteten Bereichen, mit 90 Punkten allerdings im Bereich der Mathematik in der Nähe des internationalen Mittels Streuung
- Naturwissenschaften: Standardabweichung bei 101 Punkten, damit einerseits größer als in Mathematik, anderseits deutlich über dem Wert der anderen teilnehmenden Staaten
- Streuung von Kompetenzen der fünften bis zur zehnten Klasse, sogar innerhalb einer einzelnen Schulform Leistungsunterschiede von über zwei Schuljahren
- Streuung in den Naturwissenschaften geringer als in Mathematik
- Homogene Leistungsbilder nur in Staaten, die insgesamt schwächere Leistungen erbringen (Beispiel: Portugal, Iran)

Zwar gibt es auch einige deutsche Schülerinnen und Schüler, die Spitzenleistungen im Bereich Mathematik erreichen, dies sind allerdings nur sehr wenige. International erreichen in TIMSS II 10 Prozent der Schülerinnen und Schüler einen Punktwert von über 656 Punkten. Somit müssten auch in Deutschland 10 Prozent diesen Wert erreichen oder übertreffen, es sind aber nur 6 Prozent. TIMSS III stellte fest, dass die leistungsstärksten Schülerinnen und Schüler aus Deutschland nicht mit den starken Schülerinnen und Schülern aus anderen europäischen Staaten mithalten können. Spitzenleistungen

Die Ergänzungsstudie zu TIMSS 1995 zeigte, dass Schülerinnen und Schüler aus den neuen Bundesländern bessere Ergebnisse erreichen. Diese sind wahrscheinlich darauf zurückzuführen, dass dort Mädchen besser an gymnasialer Bildung beteiligt und leistungsschwächere Schülerinnen und Schüler besser gefördert wurden. Ergänzungsstudie

Weitere Erkenntnisse aus der Ergänzungsstudie:

- Innerhalb der verschiedenen Schulformen konnte festgestellt werden, dass sich die Fähigkeiten überlappen, 30 Prozent der Realschüler und 25 Prozent der Gesamtschüler liegen mit ihren Fähigkeiten oberhalb des durchschnittlichen Niveaus der Gymnasiasten. Schulformen
- Vor allem in der Realschule ließ sich eine große Variabilität feststellen. Generell konnte in beiden Fachbereichen erkannt werden, dass die unterschiedlichen Schulformen auch unterschiedliche Ergebnisse erreichen.
- Leistungsfortschritte, die Schülerinnen und Schüler in Deutschland von der siebten zur achten Klasse erreichen, sind im internationalen Vergleich nur gering.

In allen Schulformen und in beiden untersuchten Fachrichtungen erreichen Mädchen schlechtere Leistungen als Jungen, vor allem in Physik sind die Unterschiede besonders groß (Unterschied in den neuen Bundesländern aber deutlich geringer als in den alten Bundesländern). Mädchen / Jungen

Durch die Fragebögen wurde in TIMSS II deutlich, dass die Schule bei der Identitätsbildung der Schülerinnen und Schüler behilflich sein und den Prozess des Erwachsenwerdens stabilisieren kann. Die ‚Schulunlust' bleibt innerhalb der Untersuchungen von TIMSS unverändert und unterscheidet sich auch nicht in Geschlecht oder Schulform. Das Interesse an den Gegenständen des Unterrichts nimmt innerhalb des Untersuchungszeitraumes ab. Wichtige Erkenntnis ist, dass das Selbstwertgefühl der Schülerinnen und Schüler im achten Schuljahr sinkt, das Vertrauen in die schulische Leistungsfähigkeit allerdings Fragebögen

stabil bleibt. Dabei gibt es deutliche Unterschiede zwischen Mädchen und Jungen. Selbstzweifel und Leistungsängste sind bei Mädchen verstärkt dokumentiert. Dies ist eine Erkenntnis, die international gültig ist. Dabei hängen Selbstzweifel und Leistungsängste nicht mit schlechteren Leistungen in den Fächern zusammen.

TIMSS 2007

2007 wurden in Deutschland im Rahmen von TIMSS nur Grundschülerinnen und Grundschüler am Ende der vierten Klasse. Die Ergebnisse im Vergleich zum internationalen Mittelwert zeigt folgende Tabelle 12:

	Internationaler Durchschnitt	Deutscher Mittelwert
Mathematik	473 Punkte	525 Punkte
Naturwissenschaften	476 Punkte	528 Punkte

Tabelle 12: Internationaler und deutscher Durchschnitt bei TIMSS 2007 im Vergleich

Es besteht allerdings kaum Abstand zu den anderen europäischen Staaten, dagegen gibt es große Unterschiede zu den besser abschneidenden Staaten wie beispielsweise Japan. Leider konnten keine Langzeitentwicklungen aufgezeigt werden, da Deutschland sich zum ersten Mal beteiligte.

Streuung

Zudem konnte für die deutschen Schülerinnen und Schüler eine geringere Leistungsstreuung festgestellt werden, da die Standardabweichung vom deutschen Mittelwert in Mathematik bei 68 Punkten lag. Ein solch homogenes Bild lässt sich ebenfalls für die Naturwissenschaftsleistungen feststellen.

Kompetenzniveaus Mathematik

Leider zeigten auch die Ergebnisse aus 2007 eine hohe Anzahl von Schülerinnen und Schülern mit geringen mathematischen Kompetenzen. So befindet sich etwa ein Fünftel auf den Kompetenzniveaus I und II. Zudem existiert ein Anteil von vier Prozent der Kinder, die nur mangelhafte Leistungen auf dem untersten Kompetenzniveau erreichen. Auf dem höchsten Kompetenzniveau gibt es sechs Prozent der Schülerinnen und Schüler, die mit ihren mathematischen Fähigkeiten komplexe Probleme lösen können. Diese Verteilung wird von Abbildung 5 nochmals verdeutlicht:

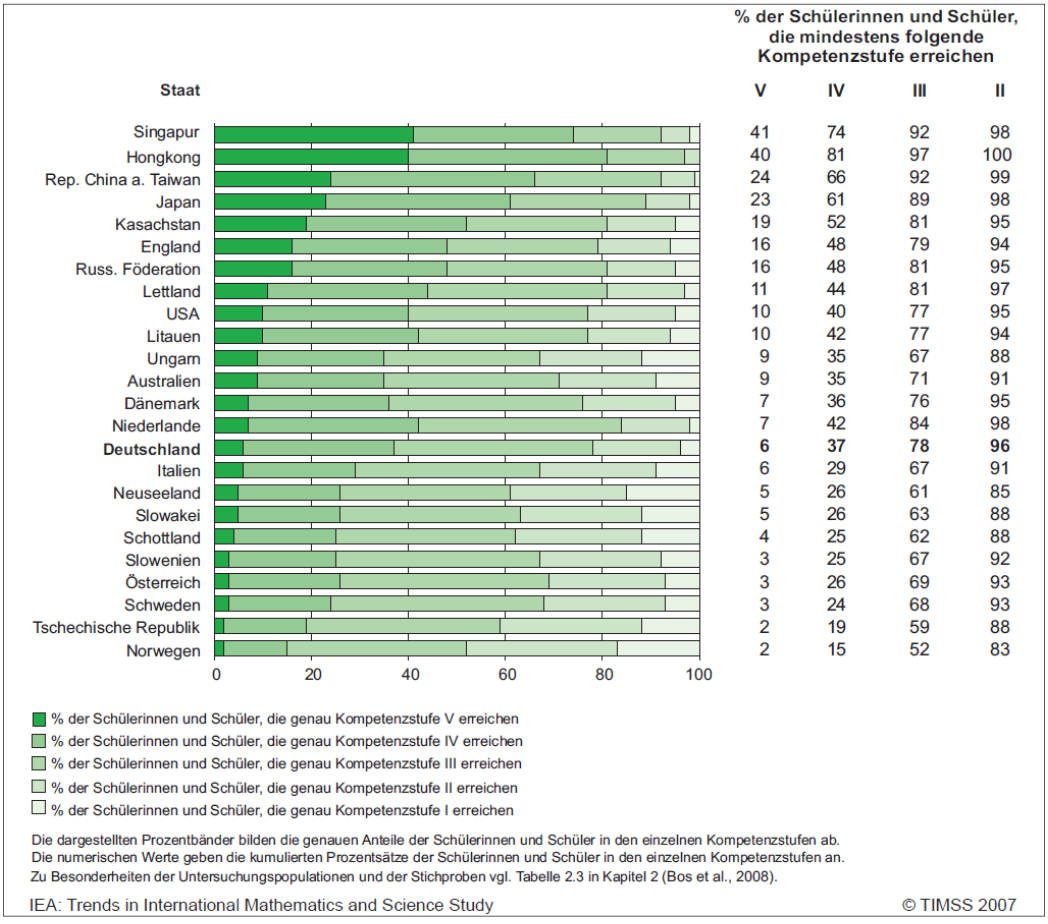


Abbildung 5: Verteilung der teilnehmenden Staaten auf die Kompetenzniveaus des Bereichs Mathematik

Naturwissenschaften

Bei den Naturwissenschaften befindet sich etwa ein Viertel der Schülerinnen und Schüler auf den beiden untersten Kompetenzniveaus, sechs Prozent erreichen nur rudimentäre Kenntnisse und liegen unter dem untersten Niveau. Das höchste Niveau wird von zehn Prozent im naturwissenschaftlichen Bereich erreicht. Sowohl die Verkleinerung des Anteils der Schülerinnen und Schüler auf der untersten Stufe als auch die Vergrößerung des Anteils der Schülerinnen und Schüler auf der obersten Stufe wird von vielen anderen Staaten, beispielsweise Japan, wesentlich besser gemeistert.

Kompetenzniveaus

Bei der Frage nach einem geschlechtsspezifischen Unterschied konnte Folgendes festgestellt werden:

Mädchen / Jungen

- Im Bereich der Mathematik lassen sich international keine geschlechtsspezifischen Unterschiede feststellen, Jungen und Mädchen erreichen beide einen Mittelwert von 473.
- In einem Drittel der teilnehmenden Staaten haben die Jungen einen deutlichen Vorsprung in Mathematik, Deutschland gehört dazu (Schüler: 531 Punkte, Schülerinnen: 519 Punkte).
- Naturwissenschaften: International zeigen sich ebenfalls keinerlei geschlechtsspezifische Unterschiede.
- Deutschland ist das einzige Land unter den EU- beziehungsweise OECD-Staaten, in dem es einen großen Unterschied in den Naturwissenschaften zwischen Mädchen und Jungen gibt

Einstellungen

- Die deutschen Grundschülerinnen und Grundschüler haben eine positive Einstellung der Mathematik gegenüber und liegen damit im internationalen Trend. Dabei ist zu erkennen, dass die Einstellung dem Fach gegenüber unabhängig von der Leistung ist. Bei 81 Prozent aller Grundschülerinnen und Grundschüler ist die Einstellung gegenüber den Naturwissenschaften positiv.

(Baumert/Lehmann, 1997; Baumert/Bos/Watermann, 1999; Köller/Baumert/Bos, 2001; Bos/Bonsen/Baumert/Prenzel/Selter/Walther, 2008)

3.4.2.2 PISA (optional)

PISA 2000

Im Jahr 2000 bildet die Lesekompetenz den Schwerpunkt der Untersuchung. Die Ergebnisse sollen stichpunkthaft zusammengefasst werden:

Lesen

- Deutschland liegt im Schwerpunktbereich Lesen mit einem Mittelwert von 484 Punkten 16 Punkte unter dem internationalen Mittelwert[18], damit gehört Deutschland zu den 14 Staaten, die sich mit ihrem Mittelwert signifikant unter dem internationalen Mittelwert befinden.

Kompetenzniveaus

- Im internationalen Vergleich ist der Anteil der Schülerinnen und Schüler, die unter dem untersten Kompetenzniveau liegen, besonders groß (international: sechs Prozent, Deutschland: zehn Prozent), viele andere Staaten (z.B. Australien und Finnland) weisen hier einen Anteil von unter fünf Prozent auf.
- In Deutschland befinden sich nur 12,7 Prozent der Schülerinnen und Schüler auf Kompetenzniveau I, fast 23 Prozent (international: 18 Prozent) liegen unter Kompetenzniveau II und können damit elementare Aufgaben nicht mit entsprechender Sicherheit lösen.
- In Deutschland liegen neun Prozent der Schülerinnen und Schüler auf dem obersten Kompetenzniveau (V) (internationaler Durchschnitt: 9,5 Prozent), damit liegt Deutschland im guten Mittelfeld. Allerdings erreichen die Spitzen wie Neuseeland und Finnland mit über 18 Prozent ein deutlich besseres Ergebnis.

Streuung

Im Vergleich zu anderen Staaten ist die Spannweite der Ergebnisse der deutschen Schülerinnen und Schüler im Bereich Lesen besonders groß. Der Leistungsabstand zwischen den fünf Prozent der besten und den fünf Prozent der schlechtesten Schülerinnen und Schüler ist im internationalen Vergleich der Größte. Hier ergibt sich ein Unterschied von 366 Punkten, der damit um 38 Punkte größer ist, als der internationale Durchschnitt der Streuung. Entsprechend diesem Befund weist Deutschland mit 111 Punkten auch die größte Standardabweichung auf.

Mathematik

Der Mittelwert von Deutschland im Bereich Mathematik liegt mit 490 Punkten zehn Punkte unter dem internationalen Mittelwert und innerhalb der unteren Hälfte des Mittelfeldes, gemeinsam mit den Vereinigten Staaten, Spanien und den osteuropäischen Staaten.

18 Unter dem internationalen Mittelwert wird im Folgenden der in den Berichten zu den Studien angegebene OECD-Mittelwert verstanden.

Kompetenzniveaus

Tabelle 13 zeigt die Verteilung der deutschen Schülerinnen und Schüler auf die Kompetenzniveaus:

Unter I	I	II	III	IV	V
7 %	17 %	32 %	31 %	12 %	1,3 %

Tabelle 13: Verteilung der deutschen Schülerinnen und Schüler auf die Kompetenzniveaus des Bereichs Mathematik bei PISA 2000

Der Anteil derjenigen Schülerinnen und Schüler, die unter Kompetenzniveau I liegen, ist in Deutschland am größten, doppelt so groß wie in Großbritannien und dreimal so groß wie in Japan. Im internationalen Vergleich setzt sich besonders Japan ab, da hier der Schwerpunkt auf den Kompetenzniveaus III bis V liegt. Hier ist der Anteil derer, die die höchste Stufe im Bereich Mathematik erreichen, sogar viermal so groß wie in Deutschland. Allerdings schneiden im oberen Bereich Staaten wie Norwegen noch schlechter ab als Deutschland.

Streuung

Die Standardabweichung liegt in Mathematik bei 103 Punkten und ist damit kleiner als im Bereich Lesen, allerdings noch immer größer als die internationale Standardabweichung von 100 Punkten. Zudem ist Deutschland auch hier wieder unter den Staaten mit der größten Streuung der Ergebnisse. Vergleicht man die Punktwerte der besten und schlechtesten fünf Prozent der Schüler, so ergibt sich ein Punktunterschied von 338 Punkten, also mehr als drei Standardabweichungen. Ähnliche Werte ergeben sich nur in wenigen weiteren Staaten.

Naturwissenschaften

Bei den Naturwissenschaften zeigen sich die folgenden Ergebnisse:

- Der Mittelwert für die deutschen Schülerinnen und Schüler liegt bei 487 Punkten und somit 13 Punkte unter dem international durchschnittlichen Mittelwert von 500 Punkten.
- Die besten Ergebnisse in diesem Bereich erreicht Korea mit 552 Punkten und liegt damit sehr weit vor Deutschland.
- Da Brasilien als schlechtestes Land nur 375 Punkte erreicht, wird deutlich, wie weit die Ergebnisse international auseinander liegen.

Kompetenzniveaus

- Die Verteilung auf die Kompetenzniveaus liegt in Deutschland anders vor, als dies international der Fall ist: Mehr Schülerinnen und Schüler liegen auf den unteren Kompetenzniveaus (52,6 Prozent auf Stufe I und II im Vergleich zu 50,1 Prozent international) und weniger auf den hohen (27,3 Prozent auf Stufe IV und V im Vergleich zu 29,8 Prozent international).

Streuung

- Die Standardabweichung liegt für die deutschen Schülerinnen und Schüler bei 102 Punkten und somit nicht weit entfernt vom internationalen Wert von 100 Punkten, gerade bei den Staaten mit sehr guten Ergebnissen ist eine wesentlich geringere Streuung der Ergebnisse vorhanden.

Ergänzungsstudie

Die Ergebnisse der Ergänzungsstudie zeigen:

- Die meisten Bundesländer liegen unter dem internationalen Durchschnitt.
- Die Länder unterscheiden sich untereinander immens, es gibt Länder, die im oberen Drittel und solche, die sich am Ende der internationalen Rangreihe befinden.

Lesen

- Für den Bereich Lesen zeigt sich in allen Bundesländern, dass die schwächsten Schülerinnen und Schüler besonders niedrige Leistungen im internationalen Vergleich erbringen, zudem haben alle Länder außer Bayern einen erhöhten Anteil von Risikoschülern, die nicht über die untersten Kompetenzniveaus hinaus kommen.

Mathematik

- Im Bereich Mathematik ergibt sich über die Bundesländer hinweg eine besonders große Streuung von 64 Punkten (Unterschied Bremen zu Bayern), zudem erreicht kein Bundesland ein Ergebnis, das mit der internationalen Spitzengruppe mithalten kann.

Naturwissenschaften

- Auch in den Naturwissenschaften ist die Streuung von 461 Punkten (Bremen) zu 508 Punkten (Bayern) sehr groß, die Leistungen innerhalb der Bundesländer unterscheiden sich also stark.
- Auch die Leistungen der besten deutschen Bundesländer sind im internationalen Vergleich nur im Mittelfeld vorzufinden.

Interessante Aspekte

Für den Bereich Lesen ergeben sich in Deutschland Hinweise auf einen Zusammenhang zwischen den Ergebnissen und der Zeit, die die Schülerinnen und Schüler freiwillig mit Lesen verbringen. In Deutschland liegt der Anteil der Schülerinnen und Schüler, die überhaupt nicht zum Vergnügen lesen, bei 42 Prozent.

Zum Thema ‚Geschlechterunterschied' lassen sich folgende Ergebnisse aufzeigen:

Geschlechterunterschied

- Der Unterschied ist international beim Lesen am größten, die mittlere Geschlechterdifferenz beträgt hier 32 Punkte, in Deutschland liegt sie mit 35 Punkten nah an diesem gemittelten Wert.
- Mädchen haben gegenüber Jungen einen Vorsprung von knapp einem halben Kompetenzniveau, allerdings gibt es Staaten, wie beispielsweise Finnland, die generell sehr gut abgeschnitten haben, wo die Geschlechterdifferenz jedoch mit 51 Punkten sehr hoch ist.
- In Mathematik zeigte sich ein Vorsprung bei den Jungen, allerdings ist dieser nicht so groß wie der der Mädchen beim Lesen (11 Punkte im internationalen Durchschnitt, 15 Punkte in Deutschland).
- Im Bereich Naturwissenschaften gibt es nur wenige Staaten, in denen sich Unterschiede zwischen Mädchen und Jungen zeigen, Deutschland gehört nicht dazu.

Zusammenhang mit sozialer Herkunft

Es wurde festgestellt, dass in keinem anderen Industrieland außer Deutschland der Zusammenhang zwischen sozialer Herkunft und fachlichen Fähigkeiten so groß ist wie in Deutschland. Vor allem Schülerinnen und Schüler mit Migrationshintergrund, in deren Familien kaum Deutsch gesprochen wird, bleiben auf den Kompetenzniveaus deutlich hinter den Anderen. Zudem konnte auch hier, wie bereits bei TIMSS, festgestellt werden, dass die Schülerinnen und Schüler, die mit 15 Jahren getestet wurden, ein bis zwei Jahre hinter Schülerinnen und Schüler anderer Staaten zurückliegen.

PISA 2003

Mathematik

Deutschland liegt beim PISA 2003-Schwerpunktthema Mathematik mit einem Mittelwert von 503 Punkten sehr nah am internationalen Mittelwert von 500 Punkten. Das beste Ergebnis im Bereich Mathematik erreicht Finnland mit 544 Punkten, ein großer Unterschied zum deutschen Ergebnis wird dabei deutlich.

Die Verteilung auf die Kompetenzniveaus zeigt Tabelle 14:

Kompetenzniveaus

Unter I	I	II	III	IV	V	VI
9,2 %	12,4 %	19 %	22,6 %	20,6 %	12,2 %	4,1 %

Tabelle 14: Verteilung der deutschen Schülerinnen und Schüler auf die Kompetenzniveaus des Bereichs Mathematik bei PISA 2003

Dabei ist zu erkennen, dass der Anteil der Schülerinnen und Schüler, die als Risikogruppe gelten, sehr groß ist, er liegt bei 21,6 Prozent. Dieser Wert ist der zweithöchste im internationalen Vergleich, der Abstand zu den anderen west- und nordeuropäischen Staaten ist sehr groß. Bei der Verteilung auf das höchste Niveau (VI) liegt Deutschland mit einem Wert von 4,1 Prozent beim internationalen Durchschnitt (4,0 Prozent). Andere Staaten (Beispiel: Belgien) erreichen hier allerdings einen Wert von 9 Prozent.

Streuung

Die Streuung fällt im internationalen Vergleich sehr unterschiedlich aus. Deutschland hat mit einer Standardabweichung von 102 Punkten den dritthöchsten Wert, nach Belgien und der Türkei.

Vergleich 2000/2003

Im Bereich Mathematik lässt sich von 2000 zu 2003 eine Verbesserung in den mathematischen Kompetenzen der Schülerinnen und Schüler in Deutschland erkennen.

Lesen

Im Bereich Lesen liegt Deutschland mit einem Mittelwert von 491 Punkten nicht weit vom internationalen Mittelwert (494 Punkte) entfernt. Auch in diesem Bereich ist der Abstand zu den besten Staaten groß, hier liegt wiederum Finnland mit 543 Punkten an der Spitze.

Streuung

Bemerkenswert ist dabei, dass gerade die Staaten mit den höchsten Ergebnissen gleichzeitig auch die niedrigsten Standardabweichungen haben. Staaten wie Finnland und Korea schaffen es, eine hohe Leistung mit geringer Streuung zu erreichen. In Deutschland liegt die Streuung mit 109 Punkten dagegen über dem internationalen Mittel von 100 Punkten.

Kompetenzniveaus

Das höchste Kompetenzniveau wird im internationalen Durchschnitt von 8,3 Prozent der Schülerinnen und Schüler erreicht. Deutschland erreicht 9,6 Prozent und liegt somit über diesem Wert, allerdings erreichen Staaten wie Neuseeland, Finnland und Australien wesentlich höhere Werte. Leider liegt Deutschland auch bei der Risikogruppe mit einem Wert von 22,3 Prozent über dem internationalen Wert (19,1 Prozent). Dieser große Anteil von Schülerinnen und Schülern, die keine einfachen Leseaufgaben bewältigen können, lag bereits im Jahr 2000 auf diesem Niveau.

Mädchen / Jungen

Dabei ist festzustellen, dass in allen teilnehmenden Staaten Mädchen bessere Leistungen erbringen als Jungen. Der Unterschied liegt im internationalen Durchschnitt bei 34 Punkten, auch in Deutschland ist er mit 42 Punkten sehr groß.

Naturwissenschaften

In den Naturwissenschaften erreichen die deutschen Schülerinnen und Schüler einen Mittelwert von 502 Punkten und befinden sich damit sehr nah am internationalen Mittelwert (500 Punkte). Damit ist eine Verbesserung von 15 Punkten gegenüber dem Ergebnis von PISA 2000 zu erkennen, und das Ergebnis liegt damit nicht mehr unter dem Durchschnitt.

Streuung

Die Streuung in den Naturwissenschaften ist in Deutschland sehr groß, sie liegt bei 111 Punkten und bildet damit, gemeinsam mit Frankreich, den höchsten Wert aller teilnehmenden Staaten. Sie hat sich gegenüber PISA 2000 noch vergrößert.

Ergäzungsstudie

Durch die Ergänzungsstudie wurde festgestellt:

- Mehrere Bundesländer haben sich im Vergleich zum internationalen Durchschnitt verbessert.
- Drei Länder liegen sogar über dem internationalen Durchschnitt.
- Obwohl eine Verbesserung deutlich erkennbar ist, zeigen in der Hälfte der teilnehmenden Bundesländer die Schülerinnen und Schüler noch immer Schwächen beim Lesen.
- Trotzdem zeigt sich, dass in vielen Bundesländern und in vielen Kompetenzbereichen eine wesentliche Verbesserung der Kompetenzniveaus stattgefunden hat, teilweise sogar mit sehr deutlichen Steigerungen.
- Steigerungen lassen sich aber nicht nur bei den Bundesländern feststellen, die schlecht abgeschnitten haben, sondern auch bei denen, die bereits gute Ergebnisse aufzuweisen hatten.

PISA 2006

Die Ergebnisse der Naturwissenschaften als Schwerpunktbereich in der Untersuchung 2006 sollen als Erstes betrachtet werden:

Naturwissenschafen

- International liegen die Mittelwerte in einem Bereich von 410 Punkten (Mexiko) bis 563 Punkten (Finnland).
- Deutsche Schülerinnen und Schüler liegen mit einem Punktwert von 516 Punkten signifikant über dem internationalen Mittelwert, somit ist ein deutlicher Zuwachs in diesem Bereich seit der Untersuchung im Jahr 2003 erkennbar.
- Trotz dieser Verbesserung gibt es noch einige Staaten, die weitaus bessere Ergebnisse vorweisen können, sodass eine Steigerung der deutschen Ergebnisse möglich ist. Gerade das Ergebnis aus Finnland zeigt, dass die Schülerinnen und Schüler dort den deutschen Schülerinnen und Schülern 1,5 bis 2 Jahre voraus sind.
- Der Anteil der Schülerinnen und Schüler, die in den Naturwissenschaften nicht über das unterste Kompetenzniveau hinaus kommen, liegt in Deutschland bei 15,4 Prozent und damit unter dem internationalen Durchschnitt, der bei 19,2 Prozent liegt.
- Spitzenleistungen: In Deutschland erreichen 11,8 Prozent der Schülerinnen und Schüler einen Punktwert, der sie auf die beiden obersten Kompetenzniveaus (V und VI) einteilt, wobei der internationale Durchschnittswert nur bei 9 Prozent liegt.
- Für die Naturwissenschaften zeigt sich, dass durchgeführte Veränderungen innerhalb der Schulen langsam greifen, allerdings immer noch ‚Luft nach oben' ist.
- Es zeigt sich, dass über die Jahre die Ergebnisse der schlechten Schülerinnen und Schüler in Deutschland in den Naturwissenschaften durchaus besser werden, die schlechtesten fünf Prozent erreichten im Jahr 2000 einen Wert von 314 Punkten, dieser Wert lag 2006 dagegen schon bei 345 Punkten.

Streuung

Die Streuung in Deutschland ist in den Naturwissenschaften mit einer Standardabweichung von 100 Punkten noch immer groß, allerdings ist auch hier eine Verbesserung erkennbar. Es gibt aber noch immer Staaten, die wesentlich homogenere Ergebnisse erreichen, wie beispielsweise Finnland.

Lesekompetenz

Bei der Betrachtung der Lesekompetenz muss bei PISA 2006 darauf geachtet werden, dass der internationale Mittelwert nicht bei 500 Punkten sondern bei 492 liegt. Die Ergebnisse waren hier:

- Deutschland liegt mit einem Mittelwert in der Lesekompetenz von 495 Punkten sehr knapp über dem internationalen Durchschnitt.

- Finnland und Korea erreichen Werte von 547 beziehungsweise 556 Punkten und schneiden deutlich besser ab als Deutschland, auch andere Staaten erreichen wesentlich bessere Ergebnisse als Deutschland. Kompetenzniveaus
- International sind es bei der Lesekompetenz 20,1 Prozent der Schülerinnen und Schüler, die sich unter oder auf dem untersten Kompetenzniveau mit ihrem Punktwert befinden, somit liegt Deutschland mit 20 Prozent sehr genau am internationalen Mittel.
- Spitzenbereich: 9,9 Prozent der deutschen Schülerinnen und Schüler erreichen das höchste Kompetenzniveau, im Vergleich zu 8,6 Prozent im internationalen Vergleich.

Die Standardabweichung reicht beim Lesen international von 81 Punkten (Finnland) bis zu 112 Punkten (Deutschland). Hier wird deutlich, dass die Streuung in Deutschland im internationalen Vergleich besonders groß ist. Spitzenleistungen können in diesem Bereich aber auch mit einer homogenen Ergebniszusammensetzung erreicht werden, wie das Beispiel Finnland zeigt.

Über die Jahre hinweg lässt sich eine Steigerung der Leistungen der deutschen Schülerinnen und Schüler beim Lesen erkennen, die allerdings nicht statistisch abgesichert werden kann. Außerdem hat sich der Abstand zu den Spitzenländern nicht verändert, er liegt noch immer bei 50 Punkten. Allerdings zeichnet sich in Deutschland eine verbesserte Verteilung auf die Kompetenzniveaus ab. So konnten die prozentualen Anteile auf den niedrigen Niveaus verringert und auf den oberen Niveaus vergrößert werden. Entwicklung über Jahre

In der Mathematik liegt der internationale Mittelwert bei 498 Punkten, die Ergebnisse stellen sich wie folgt dar: Mathematik

- 13 Staaten können den Mittelwert überschreiten, teilweise sogar deutlich, wie beispielsweise Finnland mit einem Mittelwert von 548 Punkten, zudem erreichen die asiatischen Staaten ebenfalls sehr hohe Werte. Diese Staaten erzielen Ergebnisse, die ein Kompetenzniveau über dem internationalen Mittelwert liegen
- Deutschland erreicht einen Mittelwert von 504 Punkten und liegt damit über dem internationalen Mittelwert, aber trotzdem sehr weit von den Spitzenländern entfernt
- Bei der Verteilung auf die Kompetenzniveaus schneidet Deutschland im internationalen Vergleich besser als der Durchschnitt, auf oder unter dem untersten Kompetenzniveau liegen international 21 Prozent der Schülerinnen und Schüler, in Deutschland sind es 19,9 Prozent Kompetenzniveaus
- es gibt Staaten, in denen dieser Anteil wesentlich höher ist, 56,5 Prozent der mexikanischen Schülerinnen und Schüler erreichen im Bereich Mathematik nur das unterste Kompetenzniveau beziehungsweise bleiben sogar noch darunter
- Im höchsten Kompetenzniveau (VI) schneidet Deutschland besser ab als der internationale Durchschnitt, dieser liegt bei 3,3 Prozent, in Deutschland können aber mehr Schülerinnen und Schüler (4,5 Prozent) dieses Niveau erreichen, den höchsten Wert mit 9,1 Prozent erreicht Finnland

Auch die Streuung unterscheidet sich teilweise deutlich von Land zu Land, teilweise sogar sehr deutlich. Die größte Standardabweichung und damit die größte Streuung ergibt sich in Belgien mit 106 Punkten. Deutschland liegt mit 99 Punkten noch signifikant über dem internationalen Mittel von 92 Punkten. Streuung

Bei der Betrachtung der Entwicklung über die Jahre ergeben sich für den Mathematik-Bereich sowohl im internationalen Durchschnitt als auch für Deutschland kaum Veränderungen von 2003 zu 2006, was allerdings auch bedeutet, dass die Steigerung von 2000 zu 2003 beibehalten werden konnte. Entwicklung über Jahre

Die Ergebnisse der Lesekompetenz als Schwerpunktbereich in der aktuellen Erhebung 2009 sollen als erstes beschrieben werden: PISA 2009 Lesekompetenz

- Der durchschnittliche Mittelwert aller OECD-Staaten liegt bei 493 Punkten, die Standardabweichung bei 93 Punkten. Die Verschiebung des Durchschnitts lässt sich durch die veränderte Zusammensetzung der OECD- Staaten erklären.
- Bei deutschen Schülerinnen und Schülern liegt der durchschnittliche Lesekompetenzwert bei 497 Punkten und somit im Bereich des OECD-Durchschnitts.
- In Deutschland entspricht der Anteil von Schülerinnen und Schülern mit schwachen Lesekompetenzen, also jene im unteren Kompetenzniveau, 18,5 Prozent. Der OECD-Durchschnitt liegt bei 18,8 Prozent. Somit liegt Deutschland sehr genau im internationalen Mittel.

- Spitzenbereich: Der Anteil von Schülerinnen und Schülern, die die Kompetenzstufen V und VI erreichen entspricht mit 7,6 Prozent exakt dem OECD-Durchschnitt.

Streuung

Die Streuung in Deutschland im Bereich der Lesekompetenz liegt bei 95 Punkten und unterscheidet sich somit kaum vom OECD- Durchschnittswert.

Trend

Der Trend, der schon 2003 und 2006 zu beobachten war, setzt sich auch in der PISA-Erhebung 2009 fort, nämlich eine kontinuierliche Verbesserung der Lesekompetenz. Während jedoch die Verbesserungen zwischen den Jahren 2000, 2003 und 2006 nicht statistisch abzusichern waren, liegt die Lesekompetenz 2009 nunmehr mit 497 Punkten signifikant über dem im Jahr 2000 gemessenen Wert von 484 Punkten.

Mathematik

Im Bereich Mathematik stellen sich die Ergebnisse wie folgt dar:

- Die durchschnittliche mathematische Kompetenz der OECD-Staaten liegt bei 496 Punkten.
- Deutschland liegt bei der aktuellen Erhebung mit einem mittleren Kompetenzwert von 513 über dem OECD-Mittelwert.
- Die relative Position Deutschlands hat sich im Vergleich zu seinen Nachbarstaaten gegenüber PISA 2006 verbessert.
- Zwischen den verschiedenen Bildungsgängen unterscheidet sich der Mittelwert mathematischer Kompetenz stark.
- Die mittlere mathematische Kompetenz der Schülerinnen und Schüler in Deutschland ist von PISA 2003 zu PISA 2009 signifikant angestiegen.
- Auch der Anteil der Jugendlichen unter der niedrigsten Stufe mathematischer Kompetenz ist in Deutschland von PISA 2003 zu PISA 2009 signifikant zurückgegangen.

Streuung

Die Streuung der mathematischen Kompetenz fällt in Deutschland signifikant höher aus als im OECD-Durchschnitt, sie beträgt 98 Punkte.

Natur-wissenschaften

Als Bilanz der erreichten Ergebnisse lässt sich festhalten, dass die durchschnittliche naturwissenschaftliche Kompetenz international bei 501 Punkten liegt, Deutschland erreicht folgende Werte:

- Deutsche Schülerinnen und Schüler erreichen einen Punktwert von 520 und liegen somit signifikant über dem OECD-Mittelwert.
- Spitzenbereich: Der durchschnittliche OECD- Anteil der Schülerinnen und Schüler, die die Kompetenzstufe V und VI erreichen entspricht 8,5 Prozent, in Deutschland liegt dieser Anteil bei 12,8 Prozent.
- Der prozentuale Anteil von Jugendlichen in Deutschland auf der Kompetenzstufe I und darunter ist signifikant niedriger.
- Die mittlere naturwissenschaftliche Kompetenz der Schülerinnen und Schüler in Deutschland hat sich zwischen PISA 2006 und PISA 2009 nicht signifikant verändert.
- In den verschiedenen Bildungsgängen variieren die naturwissenschaftlichen Kompetenzen der Schülerinnen und Schüler stark. Während an den Gymnasien ein durchschnittlicher Wert von 602 Punkten erreicht wird, erreichen Schülerinnen und Schüler der Hauptschulen lediglich einen Wert von 431 Punkten.

Streuung

Die Streuung der naturwissenschaftlichen Kompetenz beträgt in Deutschland 101 Punkte und ist signifikant höher als die durchschnittliche Streuung in den OECD-Staaten.

Für alle Bereiche zeichnet sich innerhalb der deutschen Bundesländer ein positiver Trend ab:

Natur-wissenschaften

- 2006 lagen bereits 13 Bundesländer signifikant über dem internationalen Mittelwert und nur ein Land darunter.
- Nicht geändert hat sich der weite Abstand der Bundesländer untereinander, zwischen dem besten Land (541 Punkte) und dem schlechtesten (485 Punkte) liegen 56 Punkte, dies entspricht dem Lernzuwachs von etwa zwei Jahren.
- Die besten Länder können auch im internationalen Vergleich gut mithalten.

- Bei der Verteilung der Kompetenzniveaus gibt es einige Länder in denen nur wenige Kinder auf oder unter der untersten Stufe liegen, in anderen Ländern ist der Anteil der Risikokinder allerdings deutlich größer.
- Auch bei der Lesekompetenz können sich die Bundesländer durchschnittlich verbessern, mehr Länder als zuvor schaffen es, oberhalb des internationalen Mittelwertes zu liegen, sind aber noch weit von den internationalen Spitzenländern entfernt. Lesen
- Der Abstand zwischen dem besten Land (Sachsen: 512 Punkte) und dem schlechtesten Land (Bremen: 474 Punkte) liegt bei 38 Punkten und hat sich damit im Vergleich zu 2003 verringert.
- Da sich der allgemeine Wert für Deutschland im Bereich Mathematik von 2003 zu 2006 nur gering verändert hat, ergeben sich auch auf der Ebene der Bundesländer nur geringe Veränderungen, auch hier ist der Abstand zur internationalen Spitze ist noch groß. Mathematik
- Das Land mit dem besten Ergebnis (Sachsen: 523 Punkte) liegt noch immer etwa 25 Punkte und damit international etwa ein Jahr hinter den besten Ländern.
- Die Distanz zwischen dem besten und dem schlechtesten Land hat sich zwar reduziert, liegt aber noch immer bei 45 Punkten.
- Auch die Streuung innerhalb der Länder ist im Mathematik-Bereich sehr groß.

Die Motivation der Mädchen ist im Bereich der Naturwissenschaften genauso wie ihre Freude an diesem Fächerkanon geringer als bei den Jungen. Dieses Ergebnis spiegelt sich jedoch nicht in den Kompetenzwerten wider. Allerdings sind es eher Jungen, die die oberen Kompetenzniveaus erreichen. Beim Lesen zeigt sich dagegen ein deutlicher Trend zum besseren Abschneiden der Mädchen. Dies ist sowohl im internationalen Mittel (38 Punkte besser) als auch in Deutschland (42 Punkte besser) der Fall. Es gibt aber auch Länder, wie beispielsweise die Niederlande, in denen der Unterschied mit 24 Punkten wesentlich kleiner ist. Überraschend dabei ist, dass der größte Unterschied mit 51 Punkten und damit einer halben Standardabweichung in Finnland vorliegt. Im Bereich Mathematik dagegen sind es die Jungen, die einen Vorsprung haben. Im internationalen Durchschnitt liegen sie 11 Punkte vor den Mädchen. Dabei ist der Unterschied besonders in Deutschland sehr groß, hier liegt er bei 20 Punkten, nur in Österreich (23 Punkte) und Japan (20 Punkte) ist er gleich groß oder größer. Motivation / Geschlechterunterschied

(Deutsches PISA-Konsortium, 2001; Moschner/Kiper/Kattmann, 2003; Deutsches PISA-Konsortium, 2002; PISA-Konsortium Deutschland, 2004; PISA-Konsortium Deutschland, 2005; PISA-Konsortium Deutschland, 2007; PISA-Konsortium Deutschland, 2008)

3.4.2.3 IGLU (optional)

IGLU 2001 machte deutlich, dass Schülerinnen und Schüler am Ende der vierten Klasse im internationalen Vergleich gut mithalten können. Daraus kann die Schlussfolgerung gezogen werden, dass die in den Sekundarstufen durch andere Studien festgestellten Mängel nicht aus der Grundschule übernommen werden, sondern erst später entstehen. IGLU 2001

Die Ergebnisse der deutschen Schülerinnen und Schüler kurz zusammengefasst:

- Der Mittelwert der deutschen Schülerinnen und Schüler liegt bei 539 Punkten und damit über dem internationalen Mittelwert von 500 Punkten.
- Deutschland befindet sich damit im oberen Leistungsdrittel und gleichauf mit den Staaten der Europäischen Union.
- Der Abstand zur Spitze ist groß, Schweden liegt mehr als eine Fünftel Standardabweichung entfernt.
- Zwei Drittel der Schülerinnen und Schüler erreichen Ergebnisse zwischen 472 und 606 Punkten, „die mittleren 50 Prozent der getesteten Jahrgangsstufe liegen überwiegend im Bereich des Kompetenzniveaus III“. Kompetenzstufen
- Nur acht Staaten haben eine geringere Standardabweichung als Deutschland (67 Punkte), somit ist Deutschland einer der wenigen Staaten mit geringer Streuung.
- In Deutschland liegen nur 1,3 Prozent der Schülerinnen und Schüler unter Kompetenzniveau I, international liegt dieser Wert bei 11,6 Prozent, in Vergleichsgruppe (VG) 2 (siehe Kapitel 3.2.1.2) bei 2,9 Prozent.
- Das höchste Kompetenzniveau (IV) wird in Deutschland von 18,1 Prozent der Schülerinnen und Schüler erreicht (international: 13,7 Prozent, VG 2: 18 Prozent).

- Einige Staaten (z. B. Frankreich), die in PISA deutlich besser abgeschnitten haben als Deutschland, liegen bei IGLU deutlich hinter Deutschland.

Ergänzungsstudie

In IGLU-E wurden 2001 die Bereiche Naturwissenschaften, Mathematik und Orthografie untersucht. Die Ergebnisse lauten:

Naturwissenschaften

Naturwissenschaften: Ergebnisse können mit denen von TIMSS verglichen werden, für die Schülerinnen und Schüler ergibt sich ein Mittelwert von 560 Punkten, dieser muss im Vergleich zu 524 Punkten als Mittelwert von TIMSS gesehen werden (TIMSS: internationaler Mittelwert 500, aber 3. und 4. Klasse, 4. Klasse 524 Punkte, 3. Klasse 473 Punkte)

- Deutschland liegt also über dem internationalen Mittelwert und im Vergleich mit den Ergebnissen von TIMSS im oberen Drittel.

Mathematik

- Mathematik: Vergleiche mit den Ergebnissen von TIMSS möglich, der Mittelwert der deutschen Schülerinnen und Schüler liegt bei 545 Punkten (internationaler Mittelwert: 529 Punkte), Deutschland liegt auch hier deutlich oberhalb des internationalen Mittelwerts.
- Ergebnisse der besten Staaten werden nicht erreicht.

Orthografie

- Orthografie: 45 Wörter diktiert, davon durchschnittlich 25,6 korrekt aufgeschrieben
- Standardabweichung bei 9,0 Wörtern
- Nur sehr wenige Kinder (3 von 2.951) schreiben alle Wörter richtig, die besten fünf Prozent der Schülerinnen und Schüler schreiben 40 Wörter richtig, die unteren 15 Prozent schreiben mehr als 30 Wörter falsch.

Auch in IGLU 2001 lässt sich ein Vorsprung der Mädchen beim Lesen erkennen. Allerdings ist dieser im internationalen Vergleich nicht so groß, da er in Deutschland bei 13 Punkten, im internationalen Vergleich dagegen bei 20 Punkten liegt.

Erkenntnisse aus Fragebögen

Zudem sollen einige Aspekte betrachtet werden, die durch die zusätzlichen Erhebungen anhand von Fragebögen erkennbar wurden und Erklärungsmöglichkeiten für die Ergebnisse bieten:

- Die Klassengröße in Deutschland liegt im internationalen Durchschnitt, allerdings gibt es weniger zusätzliche Lehrer, die in den großen Klassen eingesetzt werden.
- Die Lehrkräfte an deutschen Grundschulen sind im Vergleich älter, sie erhalten ein besonders hohes Gehalt.
- In vielen Staaten der Vergleichsgruppe 2 stehen den Schülerinnen und Schülern mehr Computer zur Verfügung, sodass dort der Computer häufiger als Arbeitsmittel eingesetzt werden kann.
- Schülerinnen und Schüler in anderen Staaten verfügen bereits bei der Einschulung über Lesevorkenntnisse, da diese dort im Rahmen des Kindergartens ausgebildet werden, in Deutschland ist dies nicht der Fall.
- Deutsche Schulleiterinnen und Schulleiter sind vor allem Lehrkräfte, dies ist in vielen anderen Staaten anders, hier sind die Schulleiterinnen und Schulleiter für die Qualität der Schule und des Unterrichts zuständig und unterrichten nur sehr wenig.

Kooperation von Lehrkräften

- In vielen Staaten werden Lehrerinnen und Lehrer dazu angehalten, untereinander zu kooperieren, dies wird durch feste Zeiten und Absprachen unterstützt, in Deutschland basiert eine Zusammenarbeit der Lehrkräfte auf Einzelinitiativen. Differenziert wird in Deutschland vor allem, um schwache Schülerinnen und Schüler zu unterstützen, dies ist in den meisten Staaten so, dazu erhalten diese Schülerinnen und Schüler in Deutschland mehr Bearbeitungszeit für die gleichen Aufgaben, dieses Verfahren wird ebenfalls von mehreren Staaten angewendet.

IGLU 2006

Bei IGLU 2006 erreicht Deutschland einen Mittelwert von 548 Punkten, liegt damit weit über dem internationalen Mittelwert (506 Punkte) und auch über den Werten der Vergleichsgruppen. Mit diesem Ergebnis liegt Deutschland im oberen Viertel aller teilnehmenden Staaten. Innerhalb der Europäischen Union gibt es kein Land, das besser abschneidet als Deutschland. Die Standardabweichung der deutschen Schülerinnen und Schüler liegt bei 67 Punkten und damit weit unter dem internationalen Durchschnitt von 103 Punkten. In der deutschen Stichprobe gibt es nur wenige Schülerinnen und Schüler, die als Risikokinder gelten, nur zwei Staaten erreichen einen geringeren Anteil. Allerdings ist der Anteil der Spitzenleser (sie erreichen das höchste Kompetenzniveau) nicht befriedigend, da er nur bei 10,8 Prozent liegt.

Kompetenzstufen

Im intranationalen Vergleich zeigen sich folgende Ergebnisse:

Bundesländer im Vergleich

- Lesekompetenz: große Unterschiede, Thüringen mit einem Mittelwert von 564 Punkten weit vor Bremen mit 522 Punkten. Thüringen erreicht damit genauso viele Punkte wie Hongkong und liegt weit vorne an der Spitze.
- Die Streuung der Bundesländer ist sehr unterschiedlich und reicht von 59 Punkten (Thüringen, Sachsen-Anhalt, Rheinland-Pfalz) bis zu 76 Punkten (Berlin).
- Die Verteilung auf die Kompetenzniveaus fällt in den Bundesländern sehr unterschiedlich aus, so gibt es Länder, die wenige Risikokinder und viele Spitzenleser haben (Thüringen: 0,4 Prozent auf Stufe I, 14,9 Prozent auf Stufe V) und Länder, die viele Risikokinder und wenige Spitzenleser haben (Berlin: 6,7 Prozent auf Stufe I, 8,6 Prozent auf Stufe V).
- Unterschied zwischen Mädchen und Jungen: sehr unterschiedliche Ergebnisse. Der Vorsprung der Mädchen reicht von -1 Punkt in Berlin (also ein Vorsprung der Jungen) bis zu +18 Punkten (Vorsprung der Mädchen) in Sachsen-Anhalt.

Ergänzende Aspekte

Auch 2006 wurden einige ergänzende Aspekte festgestellt. Ebenso wie aus der Erhebung von 2001 ergeben sich Hinweise darauf, dass sich die Vorschulzeit auszahlt: Besuchen Kinder eine vorschulische Einrichtung, so verbessert sich beispielsweise ihre Lesekompetenz. Dieses Ergebnis konnte sowohl international als auch für Deutschland erkannt werden. Zudem wurde deutlich, dass deutsche Schülerinnen und Schüler aus bildungsnahen Elternhäusern einen deutlichen Vorsprung von 67 Punkten vor Schülerinnen und Schülern aus bildungsfernen Elternhäusern haben. Dies ist ein Aspekt, der auch durch PISA festgestellt wurde. Der Vorsprung ist in Deutschland deutlich größer als das internationale Mittel von 57 Punkten. Auch in IGLU 2006 zeigte sich, dass Mädchen besser lesen als Jungen. Allerdings ist Deutschland dabei einer der sehr wenigen Staaten, in denen die Differenz mit 7 Punkten (internationales Mittel: 13 Punkte) sehr gering ausfällt. Zudem wurde gezeigt, dass Schülerinnen und Schüler mit Migrationshintergrund im internationalen Mittel schlechter abschneiden als Schülerinnen und Schüler ohne Migrationshintergrund. Dieser Unterschied konnte allerdings in Deutschland von 2001 zu 2006 um 7 Punkte verringert werden.

Bildungsnahe / bildungsferne Elternhäuser

Schülerinnen und Schüler mit Migrationshintergrund

Entwicklung 2001–2006

Abschließend kann festgehalten werden, dass Deutschland zu den elf teilnehmenden Staaten gehört, die 2006 eine signifikant bessere Leistung erreichen als 2001. In Deutschland wird eine Steigerung von 9 Punkten erreicht. Die größte Verbesserung kann in Russland (+37), Hongkong (+36) und Singapur (+30) festgestellt werden. Dabei gibt es auch Staaten, die sich innerhalb der fünf Jahre extrem verschlechtert haben. Dazu gehören Kuwait (-66), Marokko (-27) und Rumänien (-22). Zudem kann festgestellt werden, dass Deutschland näher an die besser abschneidenden Staaten herankommt und sich von den schlechter Abschneidenden absetzen kann.

(Bos u.a., 2003; Bos u.a., 2007; Bos u.a., 2008)

3.4.2.4 DESI (optional)

Die DESI-Ergebnisse von Englisch als erster Fremdsprache orientieren sich am Gemeinsamen europäischen Referenzrahmen für Sprachen (GeR), der bereits in Studienbrief-Teil ‚Vergleichsarbeiten' thematisiert wurde. Die Ergebnisse im Bereich der mündlichen Sprechfähigkeit sagen Folgendes aus:

Mündliche Sprechfähigkeit

- Zwei Drittel der deutschen Schülerinnen und Schüler erreichen am Ende der 9. Klasse das Niveau A2[19] (Erwartungshorizont für den Hauptschulabschluss), Schülerinnen und Schüler, die sich auf diesem Niveau befinden, können sich im Alltag verständlich machen.
- Ein weiteres Drittel der Schülerinnen und Schüler erreicht bereits das Niveau B1 (Erwartungshorizont für den mittleren Schulabschluss am Ende der 10. Klasse).
- Neun Prozent der Schülerinnen und Schüler können sich auf einem noch höheren Niveau verständlich machen.

19 Der Gemeinsame europäische Referenzrahmen für Sprachen ist eingeteilt in die Niveaus A1, A2, B1, B2, C1 und C2, wobei in DESI nur von A1 bis C1 getestet wurde.

Ähnliche Ergebnisse zeigen sich auch für die anderen getesteten Bereiche. Es kann zusammengefasst werden, dass der Englischunterricht, vor allem an Gymnasien, dazu führt, dass sich eine starke Leistungsspitze von 10 bis 15 Prozent bildet, die Kompetenzen besitzt, die weit über das Geforderte hinaus gehen.

Negative Aspekte

Aber auch negative Aspekte können durch DESI sichtbar werden. So haben Hauptschulen, integrierte Gesamtschulen und Schulen mit mehreren Bildungsgängen deutliche Defizite. Dies zeigt sich am Beispiel der Hauptschule darin, dass nur etwa ein Drittel der Schülerinnen und Schüler das von den Bildungsstandards aufgestellte Regelziel in Englisch erreicht.

Deutsch

Die Ergebnisse für Deutsch lauten zusammengefasst:

Schreibkompetenz

- Zwei Drittel der Schülerinnen und Schüler sind am Ende der 9. Klasse in der Lage, Briefe verständlich zu formulieren, 13 Prozent der Schülerinnen und Schüler können sich dabei stilsicher und abwechslungsreich ausdrücken.
- Drei Viertel der Schülerinnen und Schüler erkennen klare grammatische Verstöße, schwierige Verstöße werden von einem Drittel erkannt.

Lesekompetenz

- Lesekompetenz, bezogen auf literarische und Sachtexte: Fast alle Schülerinnen und Schüler erreichen mindestens das unterste Niveau.
- Ein Drittel kann zielgerichtet lesen und Informationslücken schließen.
- Übergeordnete Textstrukturen können von jedem Sechsten erkannt und mit eigenem Wissen verbunden werden.
- Etwa die Hälfte der Schülerinnen und Schüler an Haupt- und Gesamtschulen schreibt nicht angemessene Texte und erkennt keine einfachen grammatischen Fehler.
- Gymnasium: Ein Drittel der Schülerinnen und Schüler kann stilsichere und nahezu fehlerfreie Texte schreiben.

Veränderungen über ein Schuljahr

Besondere Erkenntnisse ergeben sich bei DESI daraus, dass die Schülerinnen und Schüler zu zwei Zeitpunkten getestet wurden und sich daher Veränderungen über ein Schuljahr aufzeigen lassen. Diese liegen vor allem im Englischen, aber auch in der Sprachbewusstheit des Deutschen und werden zusammengefasst in Tabelle 15 dargestellt.

Kompetenzbereich	Veränderung
Hörverstehen/Englisch	Steigerung um 27 Punkte
Umgang mit englischsprachigen Texten	Steigerung um 23 Punkte
Umgang mit Grammatik und Sprachstil/Deutsch	Zuwachs um 35 Punkte
Lesekompetenz	Kein messbarer Anstieg
Schreiben	Inhaltliche Qualität der Texte geringfügig verbessert

Tabelle 15: Veränderungen in unterschiedlichen Kompetenzbereichen über ein Schuljahr bei DESI

Zudem wird erkennbar, dass der Deutschunterricht in allen Schulformen gleich wirksam ist, da sich die Unterschiede zwischen den Schulformen innerhalb eines Schuljahres nicht verändern. Dies ist im Englischen anders, da hier vor allem die Gymnasiasten einen großen Kompetenzzuwachs zeigen.

Mädchen / Jungen

Deutlich wird durch DESI, dass die Mädchen bei sprachbezogenen Aufgaben deutlich vor den Jungen liegen. Bei DESI liegt der Vorsprung der Mädchen in Deutsch bei 41 Punkten und in Englisch bei 31 Punkten. Schülerinnen und Schüler mit nicht-deutscher Erstsprache liegen deutlich hinter denen mit Deutsch als Erstsprache zurück. Schülerinnen und Schüler, die mehrsprachig aufwachsen, liegen dagegen weniger zurück. Schülerinnen und Schülern mit nicht-deutscher Erstsprache fällt das Erlernen von Englisch leichter. Dies trifft besonders auf die Schülerinnen und Schüler zu, die mehrsprachig aufgewachsen sind.

Unterrichtsqualität

Zudem wurden auch Aspekte der Unterrichtsqualität untersucht. Der wichtigste Befund dabei ist, dass „die Lehrkraft im Durchschnitt doppelt so viel spricht wie alle Schüler zusammen“. Die Schülerinnen und Schüler haben also im Unterricht nur sehr selten die Möglichkeit, ihre Sprachkompetenzen auszuprobieren und zu erweitern.

(Klieme, 2006; DESI-Konsortium, 2008)

3.4.3 Verwendung der Ergebnisse

Wichtig bei der Darstellung und Interpretation der Ergebnisse ist, dass die Studien des Systemmonitoring vor allem als regelmäßiges Instrument gedacht sind und daher nicht nur die Ergebnisse zu einem Zeitpunkt, sondern vielmehr der Vergleich von mindestens zwei aufeinanderfolgenden Testzeitpunkten betrachtet werden muss (Wolter, 2008). Eine solche Interpretation wird in Zukunft dadurch vereinfacht, dass die Studien in einem regelmäßigen Rhythmus durchgeführt werden.

Festzuhalten ist, dass die Studien nicht zu einem rein wissenschaftlichen Zweck durchgeführt werden. Ihr Ziel ist es (vgl. Kapitel 1.2), „professionelles pädagogisches Handeln in der Unterrichtspraxis, in Schuladministration und Bildungsverwaltung zu unterstützen". Daher werden die Ergebnisse so aufbereitet, dass sie von einer breiten Mehrheit wahrgenommen und kritisch hinterfragt werden können. Nur wenn die Ergebnisse verständlich sind, können sie für das eigene Handeln berücksichtigt und übernommen werden (Klieme/Baumert, 2001). Veränderungen im Unterrichtsalltag ergeben sich einerseits aus den reflektierten Handlungen der Lehrkräfte und andererseits aus den Impulsen der Bildungspolitik und der Bildungsadministration.

Verwendung der Studien im Schulalltag

In Studien des Systemmonitorings werden Ergebnisse für die Schülerpopulation eines Landes geschätzt. Diese müssen nicht mit den Kompetenzen der Schülerinnen und Schüler an der eigenen Schule und in der eigenen Klasse übereinstimmen. Allerdings ist es in einigen Fällen möglich, Instrumente des Systemmonitorings auch in einzelnen Schulen einzusetzen, wenn das Instrument einen Aspekt untersucht, der für die Schule von besonderem Interesse ist. Die Daten der Studie können dann als ‚Normierungsstudie' benutzt und die Ergebnisse der Schule beziehungsweise Klasse können im Vergleich zu anderen Schulen verortet werden (Klieme u.a., 2007). Ob eine solche Verwendung der Ergebnisse möglich ist, muss für jede einzelne Studie geprüft werden.

Rückmeldung an teilnehmende Schulen

Eine weitere Möglichkeit, die Ergebnisse der Studien für die eigene Arbeit zu nutzen, besteht, wenn die eigene Schule beziehungsweise Klasse an der Studie teilgenommen hat. Dann werden in einigen Studien die Ergebnisse der Schule zurückgemeldet. Dabei müssen die Rückmeldungen so gestaltet werden, dass die Schule die Informationen aufgreifen und reflektieren kann. Dabei darf die Erwartung an eine solche Rückmeldung aber nicht zu groß sein. Zudem müssen die Ziele und Probleme der jeweiligen Schule beachtet werden (Klieme/Baumert, 2001).

Umgang mit Daten

Wichtig ist es auch zu beachten, dass Schulen mittlerweile eine Unmenge von Daten zur Verfügung gestellt bekommen und mit diesen auch umgehen können müssen. Dazu müssen auch die Kompetenzen innerhalb der Schule vorhanden sein, sodass die Ergebnisse interpretiert werden können und Begrifflichkeiten nicht unklar bleiben. Zudem kann es von Schule zu Schule unterschiedlich sein, wie ein bestimmter Wert interpretiert wird. Dies hängt mit den unterschiedlichen Vorstellungen und Modellen von Schulen zusammen (Rolff, 2008).

Systemmonitoring = Intervention?

Ein Problem besteht darin, dass die Erfassung von Daten im Rahmen des Systemmonitorings in den Schulen als Intervention empfunden wird. Dieses Spannungsfeld entsteht immer dann, wenn wissenschaftliche Ergebnisse innerhalb der Praxis, in diesem Fall der politischen und pädagogischen, benutzt werden sollen. „Die Handelnden in der erstgenannten Welt wissen mehr als sie können, die Handelnden in der letztgenannten können mehr als sie wissen" (Rolff, 2008). Daher genügt es nicht, wenn die Ergebnisse in ihrer wissenschaftlichen Art und Weise an die Schulen zurückgegeben werden, sie müssen aufbereitet werden. So kann es gelingen, dass die Ergebnisse nicht als Intervention sondern als „Handlungsimpulse" verstanden werden können und eine positive Entwicklung stattfinden kann (Rolff, 2008). Dies sollte das Ziel aller am Systemmonitoring beteiligten Personen sein.

3.4.4 Weiterführende Literatur

Ausführliche Informationen über die Ergebnisse der einzelnen Studien bieten die Berichte der Studien. Die Literaturangaben können den einzelnen Kapiteln entnommen werden.

3.4.5 Verständnis-Aufgaben und Diskussionspunkte

1. *Betrachten Sie die Tabelle der Kompetenzniveaus für TIMSS 1995 und erstellen Sie Interpretationen für die anderen vier Staaten.*
2. *Erstellen Sie eine Tabelle mit den Ergebnissen aller drei PISA-Erhebungen (zu einem selbst gewählten Kompetenzbereich) und machen Sie Entwicklungen über die Jahre deutlich.*

3. *Für IGLU 2001 wird angegeben, welche Erklärungsmöglichkeiten sich durch zusätzliche Erhebungen (beispielsweise nach der Klassengröße) für die Ergebnisse ergeben. Überlegen Sie, für welche Ergebnisse die Erklärungsmöglichkeiten passen könnten.*
4. *Überlegen Sie, wie es dazu kommen kann, dass sich Staaten im Verlauf von einigen wenigen Jahren stark verbessern oder verschlechtern (Beispiel IGLU 2001 zu 2006).*

3.5 Praktische Implikationen

Kapitel 3.5 beschäftigt sich abschließend mit den praktischen Implikationen. Hier wird betrachtet, welche Auswirkungen sich durch die Systemmonitoring-Studien ergeben, bezogen auf die Bildungspolitik allgemein, vor allem aber auf den Schulalltag.

In diesem Kapitel werden folgende Fragen beantwortet:

- *Wie werden die Studien von den Beteiligten rezipiert?*
- *Welche politischen Entscheidungen wurden auf Grund der Ergebnisse getroffen?*
- *Welche Auswirkungen auf die Schulpraxis ziehen die Studien-Ergebnisse nach sich?*

Wie bereits in den vorangegangen Kapiteln beschrieben, führte TIMSS 1995 dazu, dass vermehrt Aktivitäten im Bildungswesen unternommen wurden, und zwar in drei verschiedenen Bereichen:

1. Es wurden Reformmaßnahmen für den mathematisch-naturwissenschaftlichen Unterricht beschlossen.
2. Im Bildungswesen erfolgte eine Wende mit einem Fokus auf empirische Forschung.
3. Systematisches Bildungsmonitoring wurde in seiner bereits beschriebenen Form eingeführt (Klieme/Baumert, 2001).

Spätestens PISA 2000 hat dazu geführt, dass das deutsche Bildungssystem grundlegend in Frage gestellt wurde. Was daraus folgte und aus späteren Untersuchungen noch folgen wird, sind „Veränderungen im Schulsystem, in didaktisch-methodischen Orientierungen und in der Lehrerbildung“ (Eichler, 2003). Ziel all dieser Veränderungen soll es sein, „die Qualität der Bildungssysteme zu steigern und zu sichern“ (Stanat, 2008). Dieses Kapitel soll daher zeigen, welche Folgen sich aus der Einführung eines systematischen Bildungsmonitorings in Deutschland bisher ergeben haben. Dazu sollen mehrere Aspekte beschrieben werden, die als direkte Konsequenz aus den durch die Studien gewonnenen Informationen angesehen werden können. Zuerst werden allgemein bildungspolitische Veränderungen, anschließend dann ganz konkrete Entwicklungen in der Schulpraxis betrachtet.

Fehler bei der Rezeption

Zunächst soll dargestellt werden, wie die Ergebnisse von den unterschiedlichen Akteuren im Bildungsbereich aufgenommen werden. Bei der Rezeption der Ergebnisse kam es auch zu Fehlinterpretationen, sodass die Reaktionen auf die Ergebnisse nicht immer sachgemäß waren. So wurden beispielsweise die Rangskalen mehr betrachtet als die Aussagekraft der Mittelwerte und Standardabweichungen (vgl. dazu Kapitel 3.4.2.1 bis 3.4.2.4). Einige Projekte wurden allerdings auch überhastet und ohne kritische Reflexion umgesetzt, da viele beteiligte Personen davon ausgingen, dass möglichst schnell gehandelt werden muss. Beispielsweise ist dies bei der Einführung zentraler Abiturprüfungen geschehen, da sich nicht aus den Ergebnissen ablesen lässt, dass dies zu einer zwangsläufigen Verbesserung führt (Moschner/Kiper/Kattmann, 2003).

Übersetzung der Ergebnisse

Beachtet werden muss, dass die Übersetzung der Ergebnisse in tatsächliche Veränderungsvorschläge alles andere als trivial ist. Eine Schwierigkeit liegt darin, dass aus den Ergebnissen nicht direkt entnommen werden kann, an welchen Stellen Veränderungen ansetzen müssen. Zudem ist es schwer zu bestimmen, unter welchen Bedingungen sich die vorgeschlagenen Veränderungen umsetzen lassen, wie sie von der Praxis aufgenommen und welchen Effekt sie tatsächlich haben werden. Eine Gefahr besteht vor allem darin, dass die Veränderungen auch unerwünschte Nebeneffekte zur Folge haben können, vor allem dadurch, dass nur noch die leicht überprüfbaren Ziele des Bildungssystems verfolgt werden könnten (Stanat, 2008). Darüber hinaus lassen sich Fehl- oder Überinterpretationen und daraus folgende, unreflektierte Veränderungen nicht ausschließen. Dies geschieht vor allem dann, wenn aus den empirischen Daten direkte Handlungskonsequenzen für die Schulpraxis abgeleitet werden (Stanat, 2008).

Beispiel für Fehl- beziehungsweise Überinterpretationen

PISA-Befund: Die sozialen Disparitäten sind in Deutschland besonders ausgeprägt.

Begründung: Dies liegt daran, dass die Schülerinnen und Schüler schon früh auf die verschiedenen Schulformen verteilt werden.

Forderung: Reform der Schulstruktur

ABER: Kein Beleg in den PISA-Befunden dafür, dass die Ergebnisse durch die frühe Verteilung bedingt sind!

(nach: Stanat, 2008)

Um solche Fehlinterpretationen zu vermeiden, ist es ratsam, bei der Interpretation der Ergebnisse und der Ableitung von Handlungen auf mehrere Studien zurückzugreifen, da so mehr Informationen gebündelt betrachtet werden können (Stanat, 2008). Für das oben genannte Beispiel könnten beispielsweise weitere Studien zu den Themen ‚soziale Disparitäten' und ‚Schulstruktur' zu Rate gezogen werden.

Die Resultate der regelmäßig durchgeführten Studien zeigen, dass sich die Ergebnisse deutscher Schülerinnen und Schüler langsam verbessern. Das lässt darauf schließen, dass die bisherigen Veränderungen beginnen zu wirken. Ziel der Bildungspolitik muss es sein, diese positive Entwicklung weiter voran zu treiben und sich nicht auf den aktuellen Verbesserungen ‚auszuruhen', denn neben den positiven Entwicklungen zeigen die Ergebnisse der aktuellen Studien auch, dass noch immer viele Staaten deutlich bessere Ergebnisse erreichen. Hier muss vor allem erforscht werden, warum sich in anderen Staaten bessere Ergebnisse zeigen. Dies ist allerdings schwierig, da, wie bereits beschrieben, nicht festgestellt werden kann, welcher Aspekt das gute Abschneiden bedingt. So ist in Hinblick auf das obige Beispiel festzustellen, dass es sowohl Staaten ohne als auch mit gegliedertem Schulsystem gibt, die in den Studien schlecht abschneiden (Bos/Postlethwaite, 2001).

Positiver Trend

3.5.1 Rezeption der Studien (optional)

Die Ergebnisse der Systemmonitoring-Studien werden sehr unterschiedlich aufgefasst. Dabei ist leider auffällig, dass die Rezeption manchmal stark verkürzt oder sehr selektiv durchgeführt wird. In einigen Fällen werden den Studien auch Argumente entnommen, die durch die Studie selbst nicht bestätigt werden konnten (Kohler, 2005).

In diesem Kapitel soll dargestellt werden, wie die Ergebnisse der Studien von den verschiedenen Akteuren der Bildung aufgenommen werden. Weinert (2001c) verwandelt diese ernste Szene in ein ironisches Szenario und gibt ein Beispiel, wie ein Gymnasiallehrer auf die Ergebnisse von TIMSS reagiert haben könnte:

„Was mögen seine Gedanken, seine Gefühle und seine Erwartungen gewesen sein? Abschiebung jeglicher Verantwortung auf die Politik, auf die Schuladministration oder auf das mehr oder minder anonym definierte deutsche Bildungswesen? Impulsive Infragestellung der berichteten empirischen Befunde; Bedenken über die Vergleichbarkeit von Leistungswerten aus unterschiedlichen Ländern; Grimm über die Nichtberücksichtigung nachteiliger deutscher Bildungsfaktoren? Waren es persönliche Betroffenheit, Ärger, vielleicht sogar Zorn über eine neue Störung des schwierigen Dialogs zwischen Schule und Öffentlichkeit, zwischen überforderten einzelnen Lehrern und besorgten Eltern?"

Ähnliche Vermutungen lassen sich auch für die anderen Beteiligten anstellen. Dabei wird schnell nachvollziehbar, dass vor allem Bildungspolitiker dazu neigen, schnelle Lösungen finden zu wollen, um sich nicht mehr entsprechenden Fragen stellen zu müssen. Sinnvolle Reaktionen benötigen aber einen bestimmten Zeitraum um sich zu entwickeln, und bessere Ergebnisse können sich nicht von einem Tag auf den anderen einstellen (Weinert, 2001c).

Kohler (2005) hat eine Studie durchgeführt, um die Rezeption der TIMSS-Ergebnisse von TIMSS der unterschiedlichen Akteure zu untersuchen. Dabei wurden die folgenden drei Personengruppen befragt:

Studie zur Rezeption

- Lehrerinnen und Lehrer
- Eltern
- Schuladministration

Einige besonders interessante und hilfreiche Ergebnisse dieser Studie sollen hier kurz vorgestellt werden. Informationen zu Anlage und Durchführung dieser Studie können der Literatur entnommen werden.

Lehrerinnen und Lehrer

Durch Kohlers Untersuchung wurde deutlich, dass sich die Lehrerinnen und Lehrer über die Studie (in diesem Fall: TIMSS) für nicht gut informiert hielten und ihr Vorwissen darüber eher gering war. Die Einstellungen der Lehrerinnen und Lehrer zu TIMSS waren sehr unterschiedlich, sie reichten von „Was bringt so eine Studie letztendlich?“ bis zu „Man wurde über eigene Meinung durch die Studie bestätigt.“ Zudem schätzten Lehrerinnen und Lehrer, egal welcher Schulform, Fächer und Geschlecht, die Bedeutung von TIMSS gleich ein. Allerdings fanden Lehrerinnen und Lehrer mit Leitungsfunktion die Studie bedeutsamer. Auf die Frage danach, warum die deutschen Schülerinnen und Schüler nicht gut abgeschnitten haben, suchten Lehrerinnen und Lehrer die Antworten vermehrt außerhalb und nicht innerhalb der Schule. Allerdings sahen viele Lehrerinnen und Lehrer die Notwendigkeit, auf Grund der Ergebnisse etwas an ihrem Unterricht zu verändern. An die Bildungspolitik stellten sie Forderungen nach kleineren Klassen, reformierten Lehrplänen, mehr Förderunterricht und Fortbildungsmaßnahmen (Kohler, 2005).

Eltern

Auf der Seite der Eltern kann ganz allgemein davon ausgegangen werden, dass sich ihr Interesse an der Bildung ihres Kindes vergrößert hat. Sie haben bestimmte Ansprüche an die Schule und möchten mitentscheiden, wie der weitere Bildungsweg ihres Kindes gestaltet wird. Problematisch kann dabei sein, dass die Eltern jeweils nur ihr eigenes Kind im Blick haben und daher Vorstellungen entwickeln, die nicht für eine Förderung aller Schülerinnen und Schüler hilfreich sind (Kohler, 2005). Es ist wichtig, dass Eltern eine positive Einstellung den Systemmonitoring-Studien gegenüber haben, da sie im Prozess der Qualitätsentwicklung von Schulen eine wichtige Rolle spielen.

Die in der Studie untersuchten Eltern hielten sich für sehr wenig über die Studie informiert und verfügten nur über ein sehr geringes Vorwissen, deutlich geringer als das der Lehrerinnen und Lehrer. Generell fielen die Äußerungen der Eltern über das Abschneiden der deutschen Schülerinnen und Schüler wenig homogen aus. In einem mittleren Maß fanden die Eltern die Studie bedeutsam, dies kann unter anderem damit begründet werden, dass die Eltern sich mehr für das Abschneiden des eigenen Kindes interessieren. Gegenüber den Studien des Systemmonitorings waren die Eltern allgemein sehr positiv eingestellt, allerdings waren sie, ähnlich wie die Lehrerinnen und Lehrer, skeptisch, ob sinnvolle internationale Vergleiche möglich sind. Auch die Eltern suchten die Begründungen für das schlechte Abschneiden extern und nicht bei sich. Dabei wurde besonders eine generelle Unzufriedenheit mit der Bildungspolitik deutlich. Anders als die Lehrerinnen und Lehrer sahen die Eltern allerdings nur in sehr geringem Maße eine Bedeutung der Ergebnisse für die eigene Person und die eigenen Kinder (Kohler, 2005).

Unter dem Begriff der Schuladministration werden bei Kohler (2005) Beamtinnen und Beamte der Schulaufsicht verstanden. Sie haben einen großen Erfahrungshintergrund, kennen die praktische Arbeit der Lehrerinnen und Lehrer und sind innerhalb landesweiter Reformen beteiligt.

Schuladministration

Bei den Ergebnissen der Studie von Kohler finden sich viele Übereinstimmungen der Schuladministration mit den Lehrerinnen und Lehrern und den Eltern. Die Unterschiede liegen darin, dass die Schuladministration die Bedeutung der Studie als groß einschätzte und ihren Nutzen deutlich anerkannte. Zudem wünschte sie sich mehr Informationen. Dies kann daran liegen, dass Informationen über die Leistungen des gesamten Schulsystems für die Bildungsadministration auf den ersten Blick von wesentlich größerer Bedeutung sind als für die Lehrerinnen und Lehrer und die Eltern. Bei der Frage nach Handlungsmöglichkeiten hat die Schuladministration vor allem die Kompetenzen der Lehrerinnen und Lehrer im Blick. Hier wurde eine verbesserte Aus- und Fortbildung der Lehrerinnen und Lehrer gefordert. Die Aussage, die Ergebnisse seien dadurch begründet, dass zu wenig Geld für die Schulen zur Verfügung gestellt wurde, fand in der Bildungsadministration keine Unterstützung. Ebenfalls war hier kaum die Meinung vertreten, dass die schulischen Rahmenbedingungen verändert werden müssten, ein Aspekt, der bei den Eltern und vor allem bei den Lehrerinnen und Lehrern von großer Bedeutung war.

Somit wird deutlich, dass die Ergebnisse der Studien unterschiedlich aufgefasst werden können. Daher ist es für alle Beteiligten von großer Bedeutung, dass sie ihre ersten Einschätzungen überprüfen und ihre Meinungen auf einer breiten Basis an Informationen formulieren.

3.5.2 Entscheidungen auf Grund der Ergebnisse

Bereits in Kapitel 3.1 wurde beschrieben, dass die KMK nach den schlechten Ergebnissen von TIMSS handeln musste und dazu die ‚Konstanzer Beschlüsse' verfasste. Darin wurde eine regelmäßige Teilnahme an weiteren internationalen Studien beschlossen, um herauszufinden, welche Kompetenzniveaus von den deutschen Schülerinnen und Schülern erreicht werden können. Da daraufhin die deutschen Schülerinnen und Schüler auch in PISA 2000 schlecht abschnitten, wurden weitere Formen der Überprüfung wie beispielsweise Vergleichsarbeiten (vgl. Studienbrief-Teil Vergleichsarbeiten) eingeführt.

Im Jahr 2004 wurde von der KMK das ‚Institut zur Qualitätsentwicklung im Bildungswesen – Wissenschaftliche Einrichtung der Länder an der Humboldt-Universität zu Berlin' (IQB) gegründet. Eine Aufgabe des Instituts ist es, die notwendigen empirischen Testverfahren vorzubereiten (KMK, 2006). Außerdem ist es Aufgabe des Instituts, Referenzaufgaben zu entwickeln und Handbücher zu verfassen, die für die Umsetzung und Überprüfung der Bildungsstandards (vgl. Kapitel 3.5.3.1) eingesetzt werden (PISA-Konsortium Deutschland, 2007). Auf Seiten der KMK wurde zudem die Arbeit an Bildungsstandards vorangetrieben und die zentralen Abiturprüfungen entwickelt (Köller, 2008). Da sich diese Entwicklungen direkt auf den Schulalltag auswirken, werden sie genauer in Kapitel 3.5.3 beschrieben.

IQB

3.5.2.1 Sieben Handlungsfelder der KMK nach PISA 2000

Am 04.12.2001 wurden nicht nur die ersten Ergebnisse von PISA 2000 vorgestellt, sondern auch ein Handlungskatalog der KMK. Darin wurden sieben Handlungsfelder beschrieben, in denen in naher Zukunft Maßnahmen unternommen werden sollten, um auf die schlechten Ergebnisse der deutschen Schülerinnen und Schüler zu reagieren (Tillmann, 2008):

1. Verbesserung der Sprachkompetenzen in verschiedenen Bereichen
2. Bessere Verzahnung von Vor- und Grundschule; frühere Einschulung
3. Verbesserung der Grundschulbildung
4. Bessere Förderung bildungsbenachteiligter Kinder
5. Qualitätssicherung durch verbindliche Standards und Evaluation
6. Stärkung der diagnostischen und methodischen Kompetenzen der Lehrkräfte
7. Ausbau schulischer und außerschulischer Ganztagsangebote

Viele der Handlungsfelder sind komplex und beinhalten mehrere Ziele und mögliche Bereiche der Veränderung. Wie sich diese Handlungsfelder in der Praxis niedergeschlagen haben, wird Kapitel 3.5.3 zeigen. Hier werden Veränderungen in der Praxis beschrieben und mit dem jeweiligen Handlungsfeld verknüpft.

Die Beschlussfassung der Handlungsfelder führte dazu, dass die bis zu diesem Zeitpunkt teilweise sehr unterschiedlichen Aktivitäten der sechzehn Bundesländer gebündelt wurden. Bereits vor der Beschlussfassung führten einige der Bundesländer Maßnahmen in einem oder mehreren Handlungsfeldern durch. Es wird davon ausgegangen, dass diese Länder versucht haben, ihre Maßnahmen auch in der gesamtdeutschen Beschlussfassung unterzubringen, um so ihre eigene Politik abzusichern und zu legitimieren. Somit wird deutlich, dass die Handlungsfelder nicht vollkommen neue Ideen widerspiegelten, sondern bereits verfolgte Ideen zusammenfassten und für ganz Deutschland thematisierten. Das bedeutet allerdings auch, dass einige Maßnahmen unabhängig von PISA entwickelt, nach der Beschlussfassung der KMK allerdings direkt mit PISA in Verbindung gebracht wurden und dabei helfen sollten, die in PISA festgestellten Defizite zu beheben.

Bündelung der Aktivitäten

Zu beachten ist bei den Handlungsfeldern auch der politische Hintergrund, denn sie stellen nur Aspekte dar, auf die sich alle Kultusminister der Bundesländer einigen konnten. Strittige Maßnahmen, die beispielsweise nur von einer parteipolitischen Richtung vertreten wurden, wurden nicht aufgenommen, sodass beispielsweise Aussagen zum gegliederten Schulsystem komplett fehlen. Ein Vorteil dieser Verfahrensweise liegt darin, dass die Handlungsfelder Maßnahmen beschreiben, die in den Bundesländern bereits umgesetzt werden und Ergebnisse dadurch schneller sichtbar werden können. Als Nachteil erweist sich dagegen, dass, bedingt durch die zeitgleiche Darstellung der PISA-Ergebnisse, die Handlungsfelder nicht durch die Ergebnisse selbst bedingt sind, sondern bereits bestehende Entwicklungen und Maßnahmen aufnehmen (Tillmann, 2008).

3.5.2.2 Neue Schwerpunkte der KMK

Am 02.06.2006 hat die KMK eine Gesamtstrategie zum Bildungsmonitoring beschlossen. Innerhalb dieser Strategie werden vier miteinander verbundene Bereiche beschrieben, mit deren Hilfe Bildungsprozesse in Deutschland beobachtet und weiterentwickelt werden sollen. Dazu gehören:

politischer Hintergrund

- regelmäßige Teilnahme an internationalen Schulleistungsuntersuchungen,
- zentrale Überprüfung der Bildungsstandards im Ländervergleich,
- Vergleichsarbeiten (vgl. Studienbrief-Teil Vergleichsarbeiten),
- gemeinsame Bildungsberichterstattung.

Das übergeordnete Ziel dieser Gesamtstrategie liegt in der Beschaffung von Informationen, die für die Steuerung des Bildungssystems benötigt werden (vgl. Kapitel 3.1.1). Zudem sollen die so gewonnenen Erkenntnisse mit Maßnahmen der Unterrichts- und Qualitätsentwicklung verknüpft werden, sodass die pädagogische Arbeit an jeder einzelnen Schule davon profitieren kann (KMK, 2006).

Die Gesamtstrategie soll aber nicht eine umfassende Konzeption zur Weiterentwicklung des Bildungswesens darstellen. Vielmehr soll sie in bereits bestehende Beschlüsse der KMK eingebettet werden. Dazu gehören die im Folgenden beschriebenen Handlungsschwerpunkte und Arbeitsbereiche, die bereits nach PISA 2003 festgelegt wurden (KMK, 2006):

- Frühzeitige Förderung von Migranten und sozial Benachteiligten in der Bundesrepublik Deutschland
- Bereitstellung von Fortbildungskonzeptionen und -materialien zur kompetenz- bzw. standardbasierten Unterrichtsentwicklung
- Konzepte und Materialien für Deutsch als Aufgabe aller Fächer
- Aus- und Fortbildung der Lehrkräfte im Hinblick auf Verbesserung der Diagnosefähigkeit, Umgang mit Heterogenität, individuelle Förderung
- Erarbeitung eines flexibel im Unterricht einzusetzenden Aufgabenpools für die Fächer Deutsch und Mathematik

Durch die Gesamtstrategie wird deutlich, dass das Bildungsmonitoring mit konkreten Maßnahmen verknüpft werden muss. Dazu gehört neben der Unterrichts- und Qualitätsentwicklung auch die Unterstützung der Schulen (KMK, 2006).

Auf Grund der Ergebnisse von PISA 2006 und IGLU 2006 wurden von der KMK gemeinsam mit dem Bundesministerium für Bildung und Forschung im März 2008 Empfehlungen veröffentlicht. Darin werden die sieben Handlungsfelder aus dem Jahr 2001 als Grundlage für die Weiterentwicklung des Bildungssystems bestätigt und neue Schwerpunkte gesetzt (PISA-Konsortium Deutschland, 2008):

1. Stärkere Konzentration auf die Förderung leistungsschwacher Schülerinnen und Schüler insbesondere in der Sekundarstufe I
2. Durchlässigkeit des Bildungssystems verbessern, Übergänge erleichtern, Abschlüsse sichern
3. Den Unterricht weiter entwickeln, die Lehrkräfte qualifizieren

Die konkreten Veränderungen durch die Ergebnisse der Systemmonitoring-Studien und die damit verbundenen Entscheidungen zeigt das folgende Kapitel.

3.5.3 Auswirkungen des Systemmonitorings auf die Schulpraxis

Die Ergebnisse der Systemmonitoring-Studien und die sich daran anschließenden Diskussionen haben einen Einfluss auf „bildungspolitische Verläufe und schulische Innovationen“ (Tillmann, 2009) haben. Es wird durch die Studien möglich, dass bislang strittige Maßnahmen umgesetzt oder neue Pläne entwickelt werden. Dabei ist es auch möglich, dass die Ergebnisse als Legitimation für alternative Vorschläge verwendet werden, die bisher keine Mehrheit gefunden haben (Tillmann, 2009).

Als ein Beispiel für Auswirkungen des Systemmonitorings auf die Schulpraxis kann die Einführung zentraler Abiturprüfungen genannt werden. Dabei ist dies ein besonderes Beispiel, da die Einführung mit den Ergebnissen aus PISA begründet wurde, PISA selbst dazu allerdings gar keine Aussagen macht (Moschner/Kiper/Kattmann, 2003). Nähere Informationen dazu gibt Kapitel 3.5.3.4.

Eine Wirkung haben Studien im Rahmen des Systemmonitoring auf jeden Fall: Sie führen zu einer verstärkten Medienberichterstattung und erhöhen somit die Wichtigkeit der Bildungspolitik und den öffentlichen Druck auf die Bildungspolitiker.

Die folgenden Kapitel sollen in aller Kürze sechs Bereiche beschreiben, die sich durch die Systemmonitoring-Studien entwickelt beziehungsweise weiterentwickelt haben.

3.5.3.1 Standardentwicklung

Unter ‚Bildungsstandards' wird international verstanden, normative Vorgaben für die Steuerung von Bildungssystemen zu erstellen. Dabei ergeben sich unterschiedliche Arten von Standards, einige von ihnen wurden bereits in Kapitel 3.2.3.4 beschrieben.

Bereits die Handlungsfelder der KMK aus dem Jahr 2001 beinhalteten die Entwicklung von Standards. Bis zu diesem Zeitpunkt gab es in Deutschland „ein Mischsystem, in dem mit der Vorgabe von Lehrplänen, mit Kompetenzprüfungen für das Personal und mit teilstandardisierten Abgangskontrollen, z. B. im Abitur oder in der Regulierung der Anforderungen in den mittleren Abschlüssen, die Qualität der Schularbeit gesichert werden soll(te)" (Klieme u.a. 2007). Da durch PISA und TIMSS deutlich wurde, dass dieses Mischsystem nicht zum erwünschten Erfolg führte, wurde mit der Standardentwicklung begonnen.

Durch die Standards liegen ein bundesweites Referenzsystem und Bezugsgrößen für die Qualitätsentwicklung an deutschen Schulen vor (KMK, 2006). Auch wird es durch die Bildungsstandards möglich, eine Transparenz der schulischen Anforderungen herzustellen, einen kompetenzorientierten Unterricht zu entwickeln, Schülerinnen und Schüler gezielt zu fördern und erreichte Ergebnisse zu verdeutlichen (KMK, 2006). Somit gehört zur Standardentwicklung nicht nur die Erstellung der Standards, sondern auch die allgemeine Schulentwicklung und die interne und externe Evaluation (KMK, 2005). Durch die Standards wird zudem es möglich, schulische Abschlüsse unterschiedlicher Schularten zu vergleichen (KMK, 2005).

Vorteile von Bildungsstandards

Die Bildungsstandards sollen in Zukunft zeitgleich mit den Systemmonitoring-Studien überprüft werden (vgl. Kapitel 3.3.6). Dazu werden vom IQB Testverfahren entwickelt, die die Standards operationalisieren und so eine Diagnose über die erreichten Kompetenzen ermöglichen. Die Ergebnisse der Standard-Überprüfung können dann für das weitere Bildungsmonitoring verwendet werden. Zudem können so ein kriterienorientierter Vergleich durchgeführt und Aussagen über die erworbenen Kompetenzen getroffen werden. Allerdings müssen dazu hohe Qualitätsstandards, wie beispielsweise bei PISA, eingehalten werden (Klieme u.a., 2007).

Die Charakteristika von Bildungsstandards lauten (KMK, 2005):

Überprüfung der Bildungsstandards

- die Grundprinzipien des jeweiligen Unterrichtsfaches werden aufgegriffen
- die fachbezogenen Kompetenzen werden beschrieben, einschließlich zugrunde liegender Wissensbestände, die Schülerinnen und Schüler bis zu einem bestimmten Zeitpunkt ihres Bildungsganges erreicht haben sollen
- es wird auf systematisches und vernetztes Lernen abgezielt, sie folgen dem Prinzip des kumulativen Kompetenzerwerbs
- sie beschreiben erwartete Leistungen im Rahmen von Anforderungsbereichen
- sie beziehen sich auf den Kernbereich des jeweiligen Faches (Basisqualifikationen) und geben den Schulen Gestaltungsräume für ihre pädagogische Arbeit
- sie weisen ein mittleres Anforderungsniveau auf (Regelstandards)
- Aufgabenbeispiele veranschaulichen die Standards

Hinzu kommt, dass die Bildungsstandards als ‚can do statements' formuliert werden. Sie beschreiben, was ein Schüler ganz genau können soll, um den Standard zu erreichen.

Can do statements

Von Bedeutung ist dabei, dass die Standards nicht die schulischen Lehr- und Lernprozesse beschreiben oder standardisieren. Stattdessen beschreiben sie die normative Erwartung, die innerhalb der Schule erfüllt werden soll. Auf welchen Wegen dies geschieht, wie die Lernzeit eingeteilt wird und wie mit personellen Ressourcen umgegangen wird, bleibt den einzelnen Bundesländern beziehungsweise Schulen überlassen. Das bedeutet, dass nur der Output der schulischen Bildung betrachtet wird. Damit führt die Standardentwicklung auch zu einer erhöhten Eigenverantwortung der einzelnen Schulen (KMK, 2005).

Bildungsstandards = Output-Orientierung

3.5.3.2 Einführung von Kernlehrplänen

Aus der Standardentwicklung ergibt sich die Konsequenz, dass auch die Lehrpläne angepasst werden müssen. Dabei hängt die genaue Ausgestaltung der Lehrpläne davon ab, wie stark die Output-Orientierung durchgeführt wird. Wird sie radikal durchgeführt, so verlieren Lehrpläne ihre Bedeutung als „strukturierendes Element von Unterricht" (Klieme u.a., 2007). Lehrpläne, wie sie bis zum Zeitpunkt der Standardentwicklung ausgesehen haben, werden nicht mehr benötigt, wenn Standards die anzustrebenden Kompetenzen vorgeben. Bei radikaler Umsetzung kann sich jede Schule ihr eigenes Curriculum geben. In Deutschland wurde dieser radikale Weg nicht gewählt, stattdessen wurden Kernlehrpläne entwickelt.

Inhalt der Kernlehrpläne

Sie beschreiben die Standards und Erwartungen, aber auch Themen und Inhalte, gelegentlich sogar Lernformen. In Kernlehrplänen werden die Bildungsstandards mit ihrer Leitfunktion an die bisherige Orientierungsfunktion von Lehrplänen gekoppelt. Zudem soll die Autonomie jeder einzelnen Schule gestärkt werden. Ein solcher Kernlehrplan entspricht dann den nationalen Standardvorgaben, kann aber auch in zeitliche Sequenzen eingeteilt werden und direkte Empfehlungen enthalten. So schließen sich Lehrpläne und Bildungsstandards nicht gegenseitig aus, vielmehr ergänzen sie sich. Sowohl die Standards als auch die Kernlehrpläne helfen dabei, die Qualität des deutschen Bildungswesens zu steigern und eine bessere Steuerung zu ermöglichen. Während die Standards dabei nur am Output orientiert sind, betrachten die Kernlehrpläne auch den Input. Dabei liegt es im Rahmen der Möglichkeiten, dass die Festlegung von Inhalten und Lernformen immer weiter von der Landes- an die Schulebene übergeben wird (Klieme u.a., 2007).

Die Charakteristika von Kernlehrplänen lauten (Böttcher/Kalb, 2002 und Klieme u.a., 2007):

- Sie repräsentieren die Struktur von Bildung.
- Sie basieren auf Wissen (und wissensbasierten Kompetenzen).
- Sie bestimmen ein obligatorisches Fächergefüge.
- Sie nennen zentrale Themen und Inhalte, dabei aber nur das Minimum der zu Bearbeitenden.
- Sie sind offen für fachinterne Vertiefung und die thematische Koppelung von Lehrgegenständen.
- Sie bezeichnen erwartete Kompetenzen der Adressaten schulischer Arbeit. Dabei werden nur Ziele formuliert, die für alle Schülerinnen und Schüler erreichbar sind.
- Sie sind klar, eindeutig und verbindlich formuliert.
- Sie erlauben die Profilbildung von Einzelschulen.
- Sie füllen nicht mehr als 60 Prozent der Lernzeit an einer durchschnittlichen Schule.
- Sie müssen regelmäßig evaluiert und verändert werden.

Problem der Kernlehrpläne

Ein Problem der Kernlehrpläne liegt darin, dass das Bildungssystem auf ein anderes System umgestellt werden muss. Dadurch werden vor allem die Akteure im Bildungswesen vor Herausforderungen gestellt. Zunächst müssen sich daher alle an die neuen Lehrpläne gewöhnen und den Umgang mit ihnen erproben, zudem muss bei allen Akteuren eine große Akzeptanz für die Kernlehrpläne hergestellt werden.

3.5.3.3 Stärkung der Sprachkompetenz

Sprachliche Kompetenzen sind unbestritten wichtig für die Entwicklung und den beruflichen Erfolg eines Menschen. Bereits die Handlungsfelder der KMK aus dem Jahr 2001 beinhalteten daher die Förderung der Sprachkompetenz. Durch die Studien des Systemmonitorings, allen voran PISA, wurde deutlich, dass das deutsche Schulsystem nicht in der Lage ist, allen Schülerinnen und Schülern eine angemessene sprachliche Kompetenz in der Unterrichtssprache Deutsch zu vermitteln. Daher wurden verschiedene Programme aufgelegt, die die Stärkung der Sprachkompetenz zum Inhalt hatten. Dabei werden Schülerinnen und Schüler mit Migrationshintergrund, aber auch Schülerinnen und Schüler mit Deutsch als Erstsprache gefördert (Gogolin, 2008). In diesem Unterkapitel sollen beispielhaft drei Projekte zur Förderung der Sprachkompetenz dargestellt werden: FörMig, ProLesen und Delfin4, dabei wird davon ausgegangen, dass in allen Bundesländern weitere beziehungsweise ähnliche Projekte durchgeführt werden.

FörMig

Das Projekt „Förderung von Kindern und Jugendlichen mit Migrationshintergrund“ (FörMig) wurde von der Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung (BLK)[20] als große Modellinitiative als Reaktion auf das schlechte Abschneiden aller Schülerinnen und Schüler im Bereich der Lesekompetenz in PISA 2000 gestartet. Das Projekt begann Ende 2004 und endete Ende 2009. Ziel des Projektes war es, den beteiligten Bundesländern zu ermöglichen, ihre innovativen Ansätze zur Sprachförderung zu optimieren und somit für einen „Transfer guter Praxis zu sorgen“. Damit identifiziert werden kann, was genau als ‚gute Praxis‘ gelten kann, wurden Evaluationen der einzelnen Projekte durchgeführt. Zudem wurde eine zentrale Evaluation erstellt. Sie verbindet die sprachlichen Entwicklungen der beteiligten Kinder mit ihren individuellen Voraussetzungen, ihren Kontextbedingungen und den Handlungsbedingungen der beteiligten Institutionen. Die Ergebnisse der zentralen Evaluation sollen einerseits für die Bildungsplanung verwendet werden, andererseits sollen sie in die praktischen Handlungen übernommen werden. Dabei betrachtet FörMig vor allem die Übergänge von einem Bildungsweg in den anderen. Einerseits wird genau dieser Übergang untersucht und Möglichkeiten zur Gestaltung, wie

20 Auf Grund der Föderalismusreform existiert die BLK nicht mehr. Allerdings haben sich die zehn am Projekt teilnehmenden Bundesländer darauf geeinigt, das Projekt in eigener Verantwortung zu Ende zu führen.

beispielsweise Lernbiografien oder Sprachlerntagebücher, werden entwickelt, anderseits beschäftigt sich das Projekt mit der Sprachdiagnostik. Dabei ist das Projekt nicht nur auf die deutsche Sprache beschränkt, vielmehr wird davon ausgegangen, dass alle Sprachen eines Kindes wichtig für dessen weitere Entwicklung sind. Daher beinhaltet FörMig auch Einheiten, in denen die Kompetenzen in der Erstsprache der Kinder und Jugendlichen gefördert werden. Die einzelnen Projekte werden immer so konzipiert, dass Partnerschaften gebildet werden. Diese können mit Eltern, Vereinen, Bibliotheken oder auch Betrieben zustande kommen. Das abschließende Ziel des Gesamtprojektes ist es dabei nicht nur, Förderkonzepte für Deutsch zu entwickeln und zu erproben, sondern auch dazu beizutragen, dass diese Konzepte in das deutsche Bildungssystem integriert werden (Gogolin, 2008).

ProLesen

Das Projekt „ProLesen. Auf dem Weg zur Leseschule“ ist ein Projekt der KMK und basiert ebenfalls auf dem ersten KMK-Handlungsfeld aus dem Jahr 2001. Inhalt des Projektes ist die Förderung von Deutsch als Unterrichtssprache als Aufgabe aller Fächer. Dazu sollen in zehn verschiedenen Modulen für alle Schularten der Primar- und Sekundarstufen Konzepte und Materialien gesichtet, gesammelt, überarbeitet und neu entwickelt werden. Abschließend sollen diese dann aufeinander abgestimmt und zu einem Gesamtkonzept schulischer Leseförderung zusammengefasst werden. Besonders Förderkonzepte für die in PISA erkannten Risikogruppen (Lesekompetenz: Jungen und Jugendliche mit Migrationshintergrund) sollen erstellt werden. Zusätzlich soll die Diagnose- und Förderkompetenz der Lehrkräfte verbessert werden. Alle Ergebnisse des Projektes sollen gemeinsam veröffentlicht und den Schulen zugänglich gemacht werden. Abschließendes Ziel des Projekts ist es, Problembereiche zu beschreiben und Beispiele guter Praxis und Materialien zur Verfügung zu stellen. Dabei sollen die beschriebenen Maßnahmen konkret ihre Wirkungsmöglichkeiten darstellen und leicht benutzbar sein. Zusätzlich werden Beispielaufgaben angefertigt, die gleichzeitig auch der Lernstandserhebung dienen und somit den Lehrkräften bei der Diagnose behilflich sein können (Staatsinstitut für Schulqualität und Bildungsforschung München, 2008).

Delfin 4

In Nordrhein-Westfalen wird mit dem Projekt „Delfin4“ zwei Jahre vor der Einschulung die Sprachkompetenz der Kinder überprüft. Das Projekt beruht darauf, dass im Jahr 2008 die kontinuierliche sprachliche Entwicklung aller Kinder des Landes gesetzlich festgehalten wurde. Alle Kinder, auch diejenigen, die keine Kindertageseinrichtung besuchen, werden in einem zweistufigen Verfahren getestet. Dabei ist wichtig, dass die Eltern umfassend informiert werden und die Ergebnisse zeitnah erhalten. Für dieses Verfahren arbeiten die Kindertageseinrichtungen mit den nahegelegenen Grundschulen zusammen. Alle Kinder, für die ein Sprachförderbedarf festgestellt wurde, erhalten Sprachförderung. Die Teilnahme an der Sprachförderung ist verpflichtend. Das Land Nordrhein-Westfalen stellt pro an der Sprachförderung teilnehmendem Kind Geld zur Verfügung. Die Sprachförderung wird innerhalb der Kindertagesstätte durchgeführt und soll in die grundständige Sprachförderung der Einrichtung eingebettet werden. Die Ausführung und Gestaltung der Sprachförderung liegt in der Hand des Einrichtungsträgers, durchgeführt wird sie von geeigneten pädagogischen Fachkräften (Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2008).

3.5.3.4 Ganztagsschulen

Ganztagsschulen

Der Ausbau schulischer und außerschulischer Ganztagsangebote ist bereits Bestandteil der KMK-Handlungsfelder aus dem Jahr 2001. Die Aufnahme in die Handlungsfelder hat in Kombination mit den ersten PISA-Ergebnissen in mehreren Bundesländern dazu geführt, dass das Ganztags-Schulsystem intensiv ausgebaut wurde. Bis zu diesem Zeitpunkt waren die meisten deutschen Schulen Halbtagsschulen und endeten gegen 13 Uhr. In anderen Staaten wie beispielsweise England und Frankreich ist der Ganztagsschulbetrieb dagegen schon lange selbstverständlich (Tillmann, 2008).

Dabei bietet PISA keine wissenschaftlichen Ergebnisse darüber, dass die Einführung von Ganztagsschulen die Probleme des deutschen Bildungssystems lösen kann. Trotzdem waren Politiker aller Parteien der Überzeugung, dass die Einführung von Ganztagsschulen viele Probleme lösen kann. Erleichtert wurde die Einführung dadurch, dass die Bundesregierung Geld für zusätzliche 10.000 Ganztagsschulen im ganzen Land zur Verfügung stellte und so der Anreiz für die Bundesländer größer wurde, auch wenn sie selbst noch investieren mussten (Tillmann, 2009).

Im Ganztagsbereich wird durch eine Zusammenarbeit von Schule, Jugendhilfe, Kultur und Sport ein abwechslungsreiches Programm angeboten. Neben einer Hausaufgabenbetreuung werden beispielsweise

Aktivitäten im Bereich Sport und Spiel, Kreativität, Sprachförderung für Kinder mit Migrationshintergrund, Entspannung und Konzentration angeboten. Zudem erhalten die Kinder ein Mittagessen. Die Teilnahme am Ganztag ermöglicht den Kindern, sich außerhalb des Unterrichts auszuprobieren und zu erkennen, in welchen Bereichen ihre Fähigkeiten liegen. Die Schülerinnen und Schüler können somit durch ein Ganztagsangebot besser individuell gefördert werden (Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2007).

3.5.3.5 SINUS / SINUS transfer

SINUS

Das Projekt „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts“ (SINUS) wurde von der BLK auf Grund der Ergebnisse von TIMSS (vgl. Kapitel 3.4.2.1) aufgelegt. Zur Vorbereitung des Programms wurde 1997 ein Gutachten erstellt, das die Problemzonen des naturwissenschaftlichen Unterrichts beschrieb. Auf Grundlage dieses Gutachtens wurde dann die Konzeption von SINUS erstellt.

Die Konzeption des Projekts wird (nach BLK, 2001) wie folgt beschrieben:

> „Prozesse der Qualitätssicherung und Optimierung von Lehren und Lernen in den mathematisch-naturwissenschaftlichen Fächern sollen auf der Ebene der Schule in Gang gesetzt und mit dem Ziel gestützt werden, diesen eine eigene Dynamik zu geben, die über den Modellversuch hinaus trägt.“

Das Konzept beruht auf der Annahme, dass Veränderungen im Bereich Schule sich nur entwickeln und bestehen können, wenn sie von den Lehrkräften angenommen werden und diese sie in ihre Handlungsroutinen aufnehmen. Dabei setzt das Programm an den Stärken des mathematisch-naturwissenschaftlichen Themenfeldes an und schlägt Module vor, die in den Schulen und gebildeten Schulnetzen ausgewählt und bearbeitet werden können. Alle Module betreffen jeweils einen konkreten Problembereich des mathematisch-naturwissenschaftlichen Unterrichts und enthalten Hinweise auf die Möglichkeiten der Bearbeitung. Das Grundprinzip dabei ist, dass Lehrkräfte zusammenarbeiten sollen, zunächst innerhalb der Fachgruppe einer Schule, längerfristig auch über die eigene Schule hinaus.

Die einzelnen Module müssen von jeder Schule entsprechend der lokalen und regionalen Bedingungen interpretiert und konkretisiert werden.

Vor Beginn des Projekts wurden zu allen elf Modulen Handreichungen erstellt. Darin werden die Problemzonen genauer beschrieben und der Forschungsstand skizziert. Zudem werden Möglichkeiten der Bearbeitung und eventuelle Schwierigkeiten beschrieben. Beispiele sind ebenfalls enthalten. Aufgabe der Schule ist es, die Module in der praktischen Arbeit umzusetzen, dabei werden sie von wissenschaftlicher Seite unterstützt. Ihre Arbeit soll protokolliert werden, dafür stellt das Projekt Vorlagen zur Verfügung. Zudem wurde eine Internetplattform eingerichtet, auf der Informationen ausgetauscht und Kooperationen gepflegt werden können. Es ist ebenfalls Aufgabe der Schule, mit bereitgestellten Instrumenten den veränderten Unterricht zu evaluieren. Neben dieser internen Evaluation wurde das Projekt zusätzlich extern evaluiert (BLK, 2001).

Nach Ablauf des Projektzeitraums (April 1998 bis März 2001) wurde das Projekt unter dem Namen „SINUS-Transfer“ weitergeführt.

3.5.3.6 For.mat

For.mat

Das Projekt „Bereitstellung von Fortbildungskonzeptionen und -materialien zur kompetenz- bzw. standardbasierten Unterrichtsentwicklung, vor allem Lesen, Geometrie, Stochastik" ist ein Projekt der KMK, das seit Januar 2007 unter der Federführung des Landes Rheinland-Pfalz durchgeführt wurde. Innerhalb des Projekts entwickelten Personen aus allen 16 Bundesländern Materialien und Konzeptionen und stellten diese für die allgemeine Nutzung bereit. Dabei sollte ein sinnhafter Anschluss an das Projekt SINUS (vgl. Kapitel 3.5.3.5) erreicht werden. Der Unterschied zu SINUS besteht darin, dass alle Fächer betrachtet werden, für die durch die KMK Bildungsstandards beschlossen wurden. Ziel des Projekts war „die systematische Entwicklung und Qualifizierung von Fachkonferenzen und Fachgruppen zu professionellen schulinternen Lerngemeinschaften in den Bildungsstandardfächern Deutsch, erste Fremdsprache, Mathematik und Naturwissenschaften“

Das Projekt teilt sich in zwei Teilprojekte auf:

- Teilprojekt 1: Einrichtung von vier Arbeitsgruppen, diese entwickeln „fachbezogene Materialien und Konzeptionen für die kompetenzorientierte Unterrichtsentwicklung in professionellen Fachgruppen“ und stellen diese zur Verfügung
- Teilprojekt 2 hat einen „Fokus auf Kompetenzprofile und Qualifizierungskonzepte für Beraterinnen und Berater zur Unterstützung einer kompetenz- und standardbasierten Unterrichtsentwicklung“ (www.kmk-format.de, Stand 12.2.2010)

3.5.3.7 UDiKom

UDiKom

Das Projekt UDiKom, in dessen Rahmen der vorliegende Studienbrief erstellt wurde, greift das sechste Handlungsfeld der KMK von 2001 auf: „Stärkung der diagnostischen und methodischen Kompetenzen der Lehrkräfte“. Durch die Studienbrief-Module und die zusätzlichen Materialien soll vermittelt werden, wie diagnostische Instrumente für die eigene Arbeit verwendet werden können.

3.5.3.8 Bildungsberichterstattung

Bildungsberichterstattung

Die regelmäßige Bildungsberichterstattung ist Teil der KMK-Gesamtstrategie zum Bildungsmonitoring. Dabei sollen keine neuen empirischen Studien durchgeführt werden, sondern vorliegende Daten zu allen Bildungsbereichen unter der Leitidee „Bildung im Lebenslauf“ im Zusammenhang dargestellt und als Grundlage für politische Entscheidungen und für gesellschaftliche Diskussion zur Verfügung. Diese Anforderung impliziert neben der zur Verfügungstellung der Daten die Analyse, Interpretation und gesellschaftliche Einordnung. Durchgeführt wird diese Arbeit von den folgenden Institutionen: Deutsches Institut für Internationale Pädagogische Forschung (DIPF) in Kooperation mit dem Deutschem Jugendinstitut (DJI), dem Hochschul-Informations-System GmbH (HIS), dem Soziologischem Forschungsinstitut an der Universität Göttingen (SOFI) sowie den Statistischen Ämtern des Bundes und der Länder (StBA, StLÄ).

Das Ziel des Berichts liegt darin, sowohl den Zustand des deutschen Bildungssystems, die Entwicklung über die letzten Jahre als auch die aktuellen Herausforderungen zu dokumentieren. Dabei sollen die folgenden Aspekte in konzentrierter Form betrachtet werden:

- Aktuelle Situation des deutschen Bildungssystems
- Leistungsfähigkeit des deutschen Bildungssystems
- Wichtigste Problemlagen des deutschen Bildungssystems
- Bildungsprozesse im Lebenslauf
- Entwicklung des deutschen Bildungswesens im internationalen Vergleich

Der Vorteil gegenüber vielen einzelnen Berichten besteht darin, „dass die verschiedenen Bildungsbereiche in ihrem Zusammenhang gesehen, analysiert und dargestellt werden. Auf diesem Weg lassen sich übergreifende Probleme im deutschen Bildungswesen für Bildungspolitik und Öffentlichkeit sichtbar machen sowie handlungs- und steuerungsrelevante Informationen für Politik und Verwaltung gewinnen“. Der Anspruch des nationalen Bildungsberichts ist somit hoch, er will eine „Gesamtschau des Bildungssystems“ geben und greift dabei auf bereits bestehende Daten zurück, die durch „eine überschaubare Zahl von Indikatoren verdichtet“ werden. (*http://www.bildungsbericht.de/daten2008/bb_2008.pdf*, Stand 10.02.10)

Bisher wurden in den Jahren 2006 und 2008 Bildungsberichte erstellt. Die zentralen Themen, die in jedem Jahr betrachtet werden, sind:

- Bildung im Spannungsfeld veränderter Rahmenbedingungen
- Grundinformationen zu Bildung in Deutschland
- Frühkindliche Bildung, Betreuung und Erziehung
- Allgemeinbildende Schule und non-formale Lernwelten im Schulalter
- Berufliche Ausbildung
- Hochschule
- Weiterbildung und Lernen im Erwachsenenalter
- Wirkung und Erträge von Bildung

Zudem wurden in jedem Bildungsbericht Schwerpunktanalysen angestellt. Im Jahr 2006 war dies der Bereich ‚Migration', im Jahr 2008 ‚Übergänge im Anschluss an den Sekundarbereich I'.

Sehr ausführliche Informationen stehen unter www.bildungsbericht.de zur Verfügung, hier können auch die einzelnen Berichte als PDF-Dokumente heruntergeladen werden.

3.5.4 Weiterführende Literatur

Kohler (2005) beschreibt wissenschaftliche Untersuchungen zur Rezeption internationaler Schulleistungsstudien. Aktuelle Informationen über die verschiedenen vorgestellten Projekte können dem Internet entnommen werden.

FörMig: http://www.blk-foermig.uni-hamburg.de/

ProLesen: http://www.leseforum.bayern.de/index.asp?MNav=6

Delfin 4: http://www.delfin4.fb12.uni-dortmund.de/

SINUS/SINUS-Transfer: http://blk.mat.uni-bayreuth.de/indexblk.html

http://www.sinus-transfer.de/

3.5.5 Verständnis-Aufgaben und Diskussionspunkte

1. *Leiten Sie eine eigene Projektidee aus den vorgestellten Handlungsfeldern/ Schwerpunkten ab. Überprüfen Sie dabei, ob sich Ihre Idee mit einer der vorgestellten Untersuchungen legitimieren lässt.*
2. *Beschreiben Sie, analog zum Beispiel des Gymnasiallehrers, welche Fragen sich Eltern und Bildungspolitiker stellen könnten.*
3. *Überlegen Sie, in welchen Bereichen innerhalb der sieben Handlungsfelder der KMK von 2001 Veränderungsbedarf an Ihrer Bildungseinrichtung bestehen könnte.*
4. *Suchen Sie (z.B. im Internet) nach weiteren Projekten, die in den Handlungsfeldern und Schwerpunkten der KMK impliziert werden.*
5. *Beschreiben Sie, wie die acht beschriebenen Veränderungsbereiche in den Kapiteln 3.5.3.1 bis 3.5.3.8 die Ergebnisse von zukünftigen Untersuchungen verbessern können.*

3.6 Literatur

Ackeren, Isabell van (2006); *Internationale Vergleichsuntersuchungen*. Studienbrief für die Technische Universität Kaiserslautern; Fernstudium Schulmanagement, Kaiserslautern.

Albert, Ruth / Koster, Cor J. (2002); Empirie in Linguistik und Sprachlehrforschung; Gunter Narr Verlag, Tübingen

Arnold, Karl-Heinz (2001); Qualitätskriterien für die Messung von Schulleistungen; in: Weinert, Franz E. (Hg.); Leistungsmessungen in Schulen; Beltz Verlag, Weinheim u.a., S. 117-130

Autorengruppe Bildungsberichterstattung (2008); Bildung in Deutschland 2008. Ein indikatorengestützter Bericht mit einer Analyse zu Übergängen im Anschluss an den Sekundarbereich I; Bertelsmann; Bielefeld. http://www.bildungsbericht.de/daten2008/bb_2008.pdf (Stand 10.02.2010)

Baumert, Jürgen / Bos, Wilfried / Lehmann, Rainer (2000); TIMSS/III: Mathematische und naturwissenschaftliche Grundbildung am Ende der Schullaufbahn, Band 1; Leske + Budrich Verlag, Opladen

Baumert, Jürgen / Bos, Wilfried / Watermann, Rainer (1998); TIMSS/III: Schülerleistungen in Mathematik und den Naturwissenschaften am Ende der Sekundarstufe II im internationalen Vergleich ; Zusammenfassung deskriptiver Ergebnisse; Max-Planck-Institut für Bildungsforschung, Berlin

Baumert, Jürgen / Lehmann, Rainer (1997); TIMSS - mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich - deskriptive Befunde; Leske + Budrich Verlag, Opladen

Beck, Bärbel / Klieme, Eckhard (Hg.) (2007); Sprachliche Kompetenzen - Konzepte und Messung ; DESI-Studie (Deutsch Englisch Schülerleistungen International); Beltz Verlag, Weinheim u.a.

BLK – Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung (2001); Die Grundkonzeption des BLK-Programms „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts“; http://blk.mat.uni-bayreuth.de/programm/konzeption.html (Zugriff am: 30.09.2009)

Böttcher, Wolfgang / Kalb, Peter E. (2002); Kerncurriculum. Was Kinder in der Schule lernen sollen. Eine Streitschrift; Beltz Verlag, Weinheim u.a.

Bos, Wilfried / Bonsen, Martin / Baumert, Jürgen / Prenzel, Manfred / Selter, Christoph / Walther, Gerd (Hg.) (2008); TIMSS 2007 - Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich; Waxmann Verlag, Münster

Bos, Wilfried / Hornberg, Sabine / Arnold, Karl-Heinz / Faust, Gabriele / Fried, Lilian / Lankes, Eva-Maria / Schwippert, Knut / Valtin, Renate (Hg.) (2007); IGLU 2006 – Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich; Waxmann Verlag, Münster

Bos, Wilfried / Hornberg, Sabine / Arnold, Karl-Heinz / Faust, Gabriele / Fried, Lilian / Lankes, Eva-Maria / Schwippert, Knut / Valtin, Renate (Hg.) (2008); IGLU-E 2006 – Die Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich; Waxmann Verlag, Münster

Bos, Wilfried / Lankes, Eva-Maria / Prenzel, Manfred / Schwippert, Knut / Walther, Gerd / Valtin, Renate (Hg.) (2003); Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich; Waxmann Verlag, Münster

Bos, Wilfried / Postlethwaite, Thomas Neville (2001); Internationale Schulleistungsforschung: Ihre Entwicklungen und Folgen für die deutsche Bildungslandschaft; in: Weinert, Franz E. (Hg.); Leistungsmessungen in Schulen; Beltz Verlag, Weinheim u.a., S. 251-267

DESI-Konsortium (Hg.) (2008); Unterricht und Kompetenzerwerb in Deutsch und Englisch – Ergebnisse der DESI-Studie; Beltz Verlag, Weinheim u.a.

Deutsches PISA-Konsortium (Hg.) (2001); PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich; Leske + Budrich Verlag, Opladen

Deutsches PISA-Konsortium (Hg.) (2002); PISA 2000 – die Länder der Bundesrepublik Deutschland im Vergleich; Leske + Budrich Verlag, Opladen

Deutsches PISA-Konsortium (2003); PISA 2000: ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland; Leske + Budrich Verlag, Opladen

Diehl, Joerg M. / Kohr, Heinz U. (2004); Deskriptive Statistik; 13. Auflage; Verlag Dietmar Klotz, Eschborn

Döbert, Hans et al. (2009); Das Indikatoren-Konzept der nationalen Bildungsberichterstattung in Deutschland; in: Tippelt, Rudolf (Hg.); Steuerung durch Indikatoren - methodologische und theoretische Reflektionen zur deutschen und internationalen Bildungsberichterstattung; Verlag Barbara Budrich, Opladen, S. 207-272

Eichler, Wolfgang (2003); Die PISA-Nachfolgestudie DESI: „Deutsch-Englische Sprachkompetenz von Schülerinnen und Schülern der neunten Jahrgangsstufe – International“; in: Moschner, Barbara / Kiper, Hanna / Kattmann, Ulrich

(Hg.); PISA 2000 als Herausforderung. Perspektiven für Lehren und Lernen; Schneider Verlag Hohengehren, Baltmannsweiler, S. 157-171

For.Mat; Bereitstellung von Fortbildungskonzeptionen und -materialien zur kompetenz- bzw. standardbasierten Unterrichtsentwicklung, vor allem Lesen, Geometrie, Stochastik. www.kmk-format.de (Stand 12.02.2010)

Gogolin, Ingrid (2008); „Förderung von Kindern und Jugendlichen mit Migrationshintergrund FörMig" – ein länderübergreifendes Programm zur Optimierung der Sprachbildung; in: Gesellschaft. Wirtschaft. Politik (GWP); Heft 1/2008, Verlag Barbara Budrich, S. 65-75

Grotlüschen, Anke / Linde, Andrea (Hg.) (2007); Literalität, Grundbildung oder Lesekompetenz? Beiträge zu einer Theorie-Praxis-Diskussion; Waxmann Verlag, Münster

Heller, Kurt A. / Hany, Ernst A. (2001); Standardisierte Schulleistungsmessungen; in: Weinert, Franz E. (Hg.); Leistungsmessungen in Schulen; Beltz Verlag, Weinheim u.a., S. 87-101

Helmke, Andreas / Schrader, Friedrich-Wilhelm (1998). Determinanten der Schulleistung; in: Rost, Detlef H. (Hg.); Handwörterbuch Pädagogische Psychologie; Beltz Verlag, Weinheim u.a., S. 60-67

Kiper, Hanna (2003a); PISA-Ergänzungsstudie – eine erste Zusammenfassung ihrer Ergebnisse; in: Moschner, Barbara / Kiper, Hanna / Kattmann, Ulrich (Hg.); PISA 2000 als Herausforderung. Perspektiven für Lehren und Lernen; Schneider Verlag Hohengehren, Baltmannsweiler, S. 39-52

Kiper, Hanna (2003b); literacy versus Curriculum?; in: Moschner, Barbara / Kiper, Hanna / Kattmann, Ulrich (Hg.) ; PISA 2000 als Herausforderung. Perspektiven für Lehren und Lernen; Schneider Verlag Hohengehren, Baltmannsweiler, S. 65-86

Kiper, Hanna / Kattmann, Ulrich (2003); Basiskompetenzen im Vergleich – Übersicht über Ergebnisse der PISA-Studie 2000; in: Moschner, Barbara / Kiper, Hanna / Kattmann, Ulrich (Hg.); PISA 2000 als Herausforderung. Perspektiven für Lehren und Lernen; Schneider Verlag Hohengehren, Baltmannsweiler, S. 15-37

Klieme, Eckhard (2006); Zusammenfassung zentraler Ergebnisse der DESI-Studie; http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2006/2006_03_01-DESI-Ausgewaehlte-Ergebnisse.pdf (Zugriff am: 30.09.2009)

Klieme, Eckhard / Avenarius, Hermann / Blum, Werner / Döbrich, Peter / Gruber, Hans / Prenzel, Manfred / Reiss, Kristina / Riquarts, Kurt / Rost, Jürgen / Tenorth, Heinz-Elmar / Vollmer, Helmut J. (2007); Zur Entwicklung nationaler Bildungsstandards, Bundesministerium für Bildung und Forschung (Hg.), Bonn, Berlin

Klieme, Eckhard / Baumert, Jürgen (2001); TIMSS-Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente; Bundesministerium für Bildung und Forschung (Hg.); Bonn

Klieme, Eckhard / Baumert, Jürgen / Köller, Olaf (2000); Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen; in: Baumert, Jürgen / Bos, Wilfried / Lehmann, Rainer; TIMSS/III: Mathematische und naturwissenschaftliche Grundbildung am Ende der Schullaufbahn, Band 1; Leske + Budrich Verlag, Opladen, S. 85-133

Klieme, Eckhard / Neubrand, Michael / Lüdtke, Oliver (2001); Mathematische Grundbildung: Testkonzeption und Ergebnisse; in: Deutsches PISA-Konsortium (Hg.); PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich; Leske + Budrich Verlag, Opladen; S. 141-190

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2005); Bildungsstandards der Kultusministerkonferenz. Erläuterungen zur Konzeption und Entwicklung; Wolters Kluwer Verlag, München

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2006); Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring; Wolters Kluwer Verlag, München

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2009); Kompetenzstufenmodell zu den Bildungsstandards im Kompetenzbereich Sprechen und Zuhören – hier Zuhören – für den Mittleren Schulabschluss

Köller, Olaf / Baumert, Jürgen / Bos, Wilfried (2001); TIMSS – Third International Mathematics and Science Study: Dritte internationale Mathematik- und Naturwissenschaftsstudie; in: Weinert, Franz E. (Hg.); Leistungsmessungen in Schulen; Beltz Verlag, Weinheim u.a., S.269-284

Köller, Olaf (2008); Bildungsstandards in Deutschland: Implikationen für die Qualitätssicherung und Unterrichtsqualität; in: Zeitschrift für Erziehungswissenschaft; 10. Jahrgang, Sonderheft 9/2008, S. 47-59

Kohler, Britta (2005); Rezeption internationaler Schulleistungsstudien; Waxmann Verlag, Münster

Kraus, Josef (2003); TIMSS, PISA, IGLU und Co. – Fakten und Legenden; in: Kraus, Josef / Schmoll, Heike / Gauger, Jörg-Dieter; Von Timss zu IGLU. Eine Nation wird vermessen; Sankt Augustin, S. 7-72

Maritzen, Norbert (2008); Bildungsmonitoring – Systemevaluation zur Sicherung von Qualitätsstandards; in: Landesinstitut für Schule und Medien Berlin-Brandenburg (Lisum) / Kunst und Kultur (bm:ukk, Österreich) / Bundesministerium für Unterricht (Schweiz) / Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) (Hg.); Bildungsmonitoring, Vergleichsstudien und Innovationen: Von evidenzbasierter Steuerung zur Praxis; Berliner Wissenschafts-Verlag, Berlin, S. 109-124

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (2007); GanzTag, Ausgabe 01/2007, Düsseldorf

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (2008); Feststellung des Sprachstands zwei Jahre vor der Einschulung. Fachinformation zum verfahren 2009, Düsseldorf

Moschner, Barbara / Kiper, Hanna / Kattmann, Ulrich (Hg.) (2003); PISA 2000 als Herausforderung. Perspektiven für Lehren und Lernen; Schneider Verlag Hohengehren, Baltmannsweiler

OECD (2006); Assessing Scientific, Reading and Mathematical literacy – A Framework for PISA 2006; OECD-Publishing, Paris

PISA-Konsortium Deutschland (Hg.) (2004); PISA 2003 – Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs; Waxmann Verlag, Münster

PISA-Konsortium Deutschland (Hg.) (2005); PISA 2003 – Der zweite Vergleich der Länder in Deutschland – was wissen und können Jugendliche?; Waxmann Verlag, Münster

PISA-Konsortium Deutschland (Hg.) (2006); PISA 2003 – Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres; Waxmann Verlag, Münster

PISA-Konsortium Deutschland (Hg.) (2007); PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie; Waxmann Verlag, Münster

PISA-Konsortium Deutschland (Hg.) (2008); PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich; Waxmann Verlag, Münster

PISA-Konsortium Deutschland (Hg.) (2010); PISA 2009. Bilanz nach einem Jahrzehnt; Waxmann Verlag, Münster

Rauschenbach, Thomas (2009); Informelles Lernen. Möglichkeiten und Grenzen der Indikatorisierung; Tippelt, Rudolf (Hg.); Steuerung durch Indikatoren – methodologische und theoretische Reflektionen zur deutschen und internationalen Bildungsberichterstattung; Verlag Barbara Budrich, Opladen, S. 35-53

Rolff, Hans-Günter (2008); Konsequenzen aus Schulleistungsstudien und ihre Umsetzung auf Schulebene; in: Landesinstitut für Schule und Medien Berlin-Brandenburg (Lisum) / Kunst und Kultur (bm:ukk, Österreich) / Bundesministerium für Unterricht (Schweiz) / Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) (Hg.); Bildungsmonitoring, Vergleichsstudien und Innovationen: Von evidenzbasierter Steuerung zur Praxis; Berliner Wissenschafts-Verlag, Berlin, S. 147-159

Specht, Werner (2008); Innovation durch Evaluation?; in: Landesinstitut für Schule und Medien Berlin-Brandenburg (Lisum) / Kunst und Kultur (bm:ukk, Österreich) / Bundesministerium für Unterricht (Schweiz) / Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) (Hg.); Bildungsmonitoring, Vergleichsstudien und Innovationen: Von evidenzbasierter Steuerung zur Praxis; Berliner Wissenschafts-Verlag, Berlin, S. 41-52

Staatsinstitut für Schulqualität und Bildungsforschung München (2008); Projektplan ProLesen. Auf dem Weg zur Leseschule – Konzepte und Materialien zur Leseförderung als Aufgabe aller Fächer; München

Stadelmann, Willi (2008); Konsequenzen aus Schulleistungsstudien; in: Landesinstitut für Schule und Medien Berlin-Brandenburg (Lisum) / Kunst und Kultur (bm:ukk, Österreich) / Bundesministerium für Unterricht (Schweiz) / Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) (Hg.); Bildungsmonitoring, Vergleichsstudien und Innovationen: Von evidenzbasierter Steuerung zur Praxis; Berliner Wissenschafts-Verlag, Berlin, S. 169-175

Stanat, Petra (2008); Entstehung und Umsetzung von Innovationen im Bildungssystem als Konsequenz aus Bildungsmonitoring, Bildungsberichterstattung und vergleichenden Schulleistungsstudien – Möglichkeiten und Grenzen; in: Landesinstitut für Schule und Medien Berlin-Brandenburg (Lisum) / Kunst und Kultur (bm:ukk, Österreich) / Bundesministerium für Unterricht (Schweiz) / Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) (Hg.); Bildungsmonitoring, Vergleichsstudien und Innovationen: Von evidenzbasierter Steuerung zur Praxis; Berliner Wissenschafts-Verlag, Berlin, S. 11-23

ThILLM – Thüringer Institut für Lehrerfortbildung, Lehrplanentwicklung und Medien (ThILLM); Gehirngerechtes Klassenzimmer – „Handreichungen für die Unterrichtspraxis"; ThILLM-Heft 126

Tillmann, Klaus-Jürgen (2008) (Hg.); PISA als bildungspolitisches Ereignis. Fallstudien in vier Bundesländern; Verlag für Sozialwissenschaften, Wiesbaden

Tillmann, Klaus-Jürgen (2009); Was leistet die PISA-Studie zur Steuerung des Bildungssystems?; in: Tippelt, Rudolf (Hg.); Steuerung durch Indikatoren – methodologische und theoretische Reflektionen zur deutschen und internationalen Bildungsberichterstattung; Verlag Barbara Budrich, Opladen, S.17-30

Weinert, Franz E. (2001a); Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit; in: Weinert, Franz E. (Hg.); Leistungsmessungen in Schulen; Beltz Verlag, Weinheim u.a., S. 17-31

Weinert, Franz E. (Hg.) (2001b); Leistungsmessungen in Schulen; Beltz Verlag, Weinheim u.a.

Weinert, Franz E. (2001c); Perspektiven der Schulleistungsmessung – mehrperspektivisch betrachtet; in: Weinert, Franz E. (Hg.); Leistungsmessungen in Schulen; Beltz Verlag, Weinheim u.a., S. 353-365

Wolter, Andrä (2009); Hochschulindikatoren in der nationalen Bildungsberichterstattung: Ihre Stärken und Schwächen; in: Tippelt, Rudolf (Hg.); Steuerung durch Indikatoren – methodologische und theoretische Reflektionen zur deutschen und internationalen Bildungsberichterstattung; Verlag Barbara Budrich, Opladen, S. 73-91

Wolter, Stefan C. (2008); Bildungsberichterstattung auf der Basis von Indikatoren; in: Landesinstitut für Schule und Medien Berlin-Brandenburg (Lisum) / Kunst und Kultur (bm:ukk, Österreich) / Bundesministerium für Unterricht (Schweiz) / Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) (Hg.); Bildungsmonitoring, Vergleichsstudien und Innovationen: Von evidenzbasierter Steuerung zur Praxis; Berliner Wissenschafts-Verlag, Berlin, S. 53-70

Unterrichtsdiagnostik mit EMU

Andreas Helmke
Tuyet Helmke
Gerlinde Lenske
Giang Hong Pham
Anna-Katharina Praetorius
Friedrich-Wilhelm Schrader
Manuel Ade-Thurow

UDiKom

Aus- und Fortbildung der Lehrkräfte in Hinblick auf Verbesserung der Diagnosefähigkeit, Umgang mit Heterogenität, individuelle Förderung

Unterrichtsdiagnostik mit EMU

Alle im Projekt erstellten Materialien
finden Sie unter
www.unterrichtsdiagnostik.info

4. Einführung

Ziel des Moduls *Unterrichtsdiagnostik* ist es, Lehrkräften, Referendaren/Lehramtsanwärtern und Lehramtsstudierenden Werkzeuge an die Hand zu geben, die eine kriteriengeleitete Reflexion des eigenen Unterrichts ermöglichen. Damit soll eine Grundlage für gezielte Verbesserungen des Unterrichts geschaffen werden. Wie vor allem Hattie (2013) gezeigt hat, spielt Feedback eine zentrale Rolle, wenn Veränderungen erreicht werden sollen. Fundierte Rückmeldung über den Unterricht und seine Wirkungen zu bekommen und diese zu nutzen ist keineswegs trivial. Hierzu bedarf es geeigneter Methoden und Werkzeuge. Hattie zufolge geht es darum, das Lehren und Lernen dadurch sichtbar zu machen, dass im Klassenzimmer verstärkt Situationen geschaffen werden, die eine solche Rückmeldung erlauben. Rückmeldungen bilden den Kern der Unterrichtsdiagnostik. Der springende Punkt ist dabei, sich auf ein empirisches Fundament zu stützen, also datenbasiert vorzugehen: „Hier diskutieren, bewerten und planen sie [*die Lehrer*] ihren Unterricht im Licht der Feedback-Evidenz: über den Erfolg und die weiteren Wirkungen ihrer Lehrstrategien und Konzepte, über Fortschritt und angemessene Herausforderungen. Dies ist nicht (nur) kritische Reflexion, sondern *kritische Reflexion im Licht der Evidenz,* also im Licht empirischer Belege zu ihrem Unterricht" (Hattie, 2013, S. 281).

Die bundesweite Erprobung von EMU in Schulen und Studienseminaren hat gezeigt: Für den hauptsächlichen Zweck des Moduls – die Handlungsfähigkeit von Lehrpersonen im Alltag durch eine evidenzbasierte Diagnostik des Unterrichts zu verbessern – ist es nicht zielführend, den Weg einer reinen Wissensvermittlung zu gehen, etwa in Form eines umfangreichen „Studienbriefes". Aus vielfältigen Rückmeldungen „vor Ort" haben wir einige Lektionen gelernt:

1) Nur ein handlungsorientiertes Werkzeug mit praxistauglichen Formaten hat eine Chance, akzeptiert und in der Praxis genutzt zu werden. Unser anfänglicher, ca. 130 Seiten starker Studienbrief stieß lediglich in Universitäten / Pädagogischen Hochschulen auf Resonanz, von der Schulpraxis wurde er jedoch wegen seines Umfanges stark kritisiert und als nicht praxisgerecht zurückgewiesen. Wir haben daraus gelernt und Texte entwickelt, die auf typische „Zeitfenster" in der Praxis abgestimmt sind:

- eine 1-seitige Informationsbroschüre für Personen, die sich in einer *großen Pause* über EMU informieren möchten
- eine 10-seitige Broschüre für diejenigen, die sich in einer *Freistunde* über EMU informieren möchten; für gewünschte Vertiefungen haben wir entsprechende Verweise (Hyperlinks)[1] vorgesehen. Um den Umfang der Broschüre so gering wie möglich zu halten, haben wir weiterführende Informationen sowie die Fragebögen und das EDV-Tool als Hyperlinks in die o. g. Broschüre integriert. Diese Broschüre ist am Ende dieses Kapitels abgedruckt.

2) Für die Akzeptanz in der Praxis ist es offenbar auch wichtig,

- das Angebot nach Art eines Baukastens zu flexibilisieren, um Lehrkräften je nach Bedarf unterschiedliche, insbesondere auch niederschwellige Einstiege zu ermöglichen (Modulprinzip);
- es zu ermöglichen, die Itempools der Unterrichtsbeobachtungsbögen je nach Bedarf zu ändern, eine Auswahl daraus zu treffen oder mit eigenen Items zu ergänzen, um so den Bedürfnissen der individuellen Lehrpersonen oder auch der gesamten „lernenden Schule" Rechnung zu tragen.

Die von uns entwickelten Materialien umfassen Fragebögen zur Unterrichtsbeurteilung aus verschiedenen Perspektiven (unterrichtende Lehrperson, hospitierende Lehrperson, Schülerinnen und Schüler), die oben erwähnten Broschüren, Power-Point-Folien, eine Übersicht über Unterrichtsvideos sowie ein EDV-Tool zur Eingabe, Auswertung und Visualisierung der erhobenen Daten.

Dieses Konzept war erfolgreich: Die EMU-Materialien sind seit der Freischaltung Ende Januar 2011 bereits über 150.000mal heruntergeladen worden, mit jährlich steigender Tendenz. Aktuell verfolgen wir die internationale Nutzung von EMU mit Hilfe einer Open-Source-Webanalyseplattform namens PIWIK (http://de.piwik.org/).

Zudem ist EMU inzwischen in mehreren Bundesländern bereits in das Angebot der Lehreraus- und -fortbildung integriert, so beispielsweise in Bayern, Rheinland-Pfalz, Sachsen-Anhalt, Nordrhein-Westfalen und Baden-Württemberg; desgleichen in der Schweiz (z. B. in den Kantonen Zürich, Luzern und Graubünden).

In Zusammenhang mit EMU sind zur Thematik Unterrichtsdiagnostik zahlreiche Publikationen entstanden, welche am Ende des Beitrags aufgeführt sind.

1 Die Verweise sind in der digitalen Version unter www.unterrichtsdiagnostik.info zu finden.

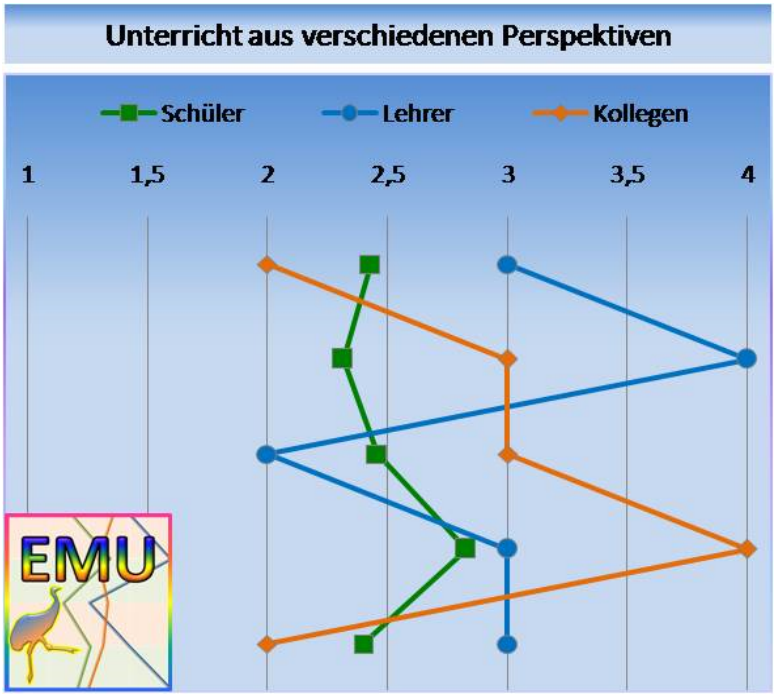
A. Helmke, T. Helmke, G. Lenske, G. Pham, A.-K. Praetorius,
F.-W. Schrader & M. Ade-Thurow

EMU

Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung

Version 4.2 (01.02.2014)

herunterladbar unter www.unterrichtsdiagnostik.info



„Der wichtigste Aspekt besteht darin, im Klassenzimmer Situationen zu schaffen, in denen die Lehrpersonen mehr Feedback über ihren Unterrichtsstil erhalten können.“ (Hattie, 2013, S. 15)

EMU steht für *Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung.* Es handelt sich dabei um ein Programm, das wir 2011 im Rahmen des Projektes UdiKom im Auftrag der Kultusministerkonferenz entwickelt haben. Weil bei EMU sicher jeder an die gleichnamige Vogelart denkt, haben wir dieses possierliche Tier in unser Logo aufgenommen.

Das auf www.unterrichtsdiagnostik.de frei verfügbare Material umfasst

- die vorliegende Broschüre sowie weiterführende Texte und Verweise auf Internetseiten
- Instrumente für die Unterrichtsbeobachtung
- Software für die Visualisierung der Ergebnisse
- PowerPoint-Folien für Einführungsveranstaltungen zur Unterrichtsdiagnostik
- Videografierte Unterrichtsstunden für Übungszwecke

Übersicht

Hier finden Sie nähere Informationen zum Autorenteam und zur Vorgeschichte von EMU

Hier geht‘s direkt zu den Fragebögen

Hier geht’s direkt zu den Auswertungsprogrammen und den Manualen

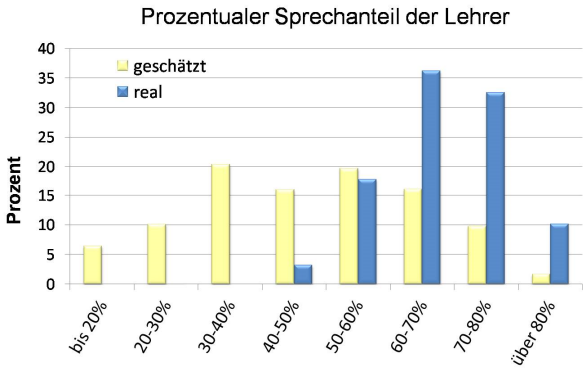
Literatur

Diese Broschüre, das zugehörige Material und die Software werden fortlaufend verbessert und ergänzt. Hinweise, Vorschläge und Fragen bitte an unterrichtsdiagnostik@gmail.com

4.1 Unterrichtsdiagnostik – was ist das, und warum ist sie nötig?

Bei *Diagnose* denkt man im alltäglichen Sprachgebrauch häufig zunächst an die Medizin. Ursprünglich bedeutet das aus dem Griechischen stammende Wort Diagnose aber einfach nur die Erforschung eines Sachverhaltes mit dem Ziel, beobachtete Merkmale einem Klassifikationssystem zuzuordnen; wer dies kann, ist ein *diagnostikos* (zum Unterscheiden begabt). Auf den Bereich des Unterrichts übertragen, heißt Diagnostik: Bestandsaufnahme. Man spricht auch von daten- oder evidenzbasiertem Vorgehen.

Im Unterricht finden solche Bestandsaufnahmen hauptsächlich in Form von offiziellen Unterrichtsbesuchen durch die Schulleitung, Lehrproben und Unterrichtsbeobachtungen im Rahmen der externen Evaluation statt. Dies sind seltene Ereignisse, die nicht immer diagnostischen Anforderungen genügen und oft mit einem Evaluationsdruck verbunden sind. Im Schulalltag bildet sich der einzelne Lehrer in der Regel nur aufgrund von unsystematischen Beobachtungen und Rückmeldungen ein Urteil über die Qualität des eigenen Unterrichts (Schrader & Helmke, 2001).



Dass das damit verbundene Wissen begrenzt ist, zeigte sich z.B. in der DESI-Videostudie des Englischunterrichts in der 9. Jahrgangsstufe (T. Helmke et al., 2008). Die nebenstehende Abbildung zeigt: Lehrpersonen halten sich für wesentlich schweigsamer und zurückhaltender als sie es (gemessen an der Echtzeitmessung auf der Grundlage einer Videoaufzeichnung) tatsächlich sind. Der eigene Sprechanteil – ein wichtiger Indikator eines schüleraktivierenden Unterrichts – wird deutlich unterschätzt.

Hierzu Hattie: *„Teachers talk, talk, and talk … Classrooms are dominated by teacher talk, and one of the themes of Visible Learning is that the proportion of talk to listening needs to change to far less talk and much more listening"* (2012, S. 72).

Ergebnisse wie dieses sind nur auf den ersten Blick erstaunlich. Eine realistische Selbsteinschätzung würde ja voraussetzen, dass man unterrichtet und gleichzeitig eine Meta-Perspektive einnimmt, von der aus man das eigene Verhalten und dessen Auswirkungen kontinuierlich beobachtet und bilanziert („monitoring"). Damit wären Lehrer angesichts der Komplexität des Lehr-Lern-Geschehens im Klassenzimmer – Multidimensionalität, Gleichzeitigkeit, Unvorhersehbarkeit, Unaufschiebbarkeit, Relevanz für künftiges Handeln (Doyle, 2006) – jedoch überfordert.

Will man den eigenen Unterricht weiterentwickeln, dann ist es zunächst einmal nötig, über zutreffende Informationen zu verfügen. Wie das Forschungsbeispiel aus der DESI-Studie zeigt, ist es dazu nötig, die eigene Sichtweise durch andere Perspektiven zu ergänzen, etwa durch kollegiale Hospitation und Schülerfeedback. Ohne einen solchen Blick von außen sind Versuche der Unterrichtsveränderung in der Gefahr des Stocherns im Nebel. Noch wahrscheinlicher ist aber, dass Unterrichtsveränderungen gar nicht erst erwogen werden, weil überhaupt kein erkannt wird.

> Je länger Lehrkräfte im Beruf sind, desto schwieriger wird es, eingefahrenen Routinen zu entkommen (...) Mit der Zeit können sich die immer gleichen ‚Fehler' einschleichen, die nicht einmal von einem selbst bemerkt werden. Wenn viele Lehrkräfte diese blinden Flecken zwar unbewusst spüren, sie aber nicht bewusst wahrnehmen und somit auch nicht ändern können, hilft hier Rückspiegelung (Feedback) durch Dritte (Horster & Rolff, 2006, S. 202f.)

4.2 An wen richtet sich EMU?

EMU richtet sich an alle, die ihren Unterricht weiterentwickeln möchten oder andere dabei beraten. Dies sind primär Lehrende und Lernende im Bereich von Schule und Lehrerausbildung, aber auch die Schulaufsicht (Zielvereinbarungen!). Die Ziele von EMU sind vielfältig:

- Erkennen von Stärken und Schwächen des eigenen Unterrichts
- Datenbasierte Weiterentwicklung des Unterrichts

- Sensibilisierung für Heterogenität in der Klasse
- Bewusstmachung eigener subjektiver Theorien des Lehrens und Lernens
- Verständigung über ein gemeinsames Bild von Unterricht im Team / Kollegium
- Schulentwicklung: Kollegialer Austausch und „Öffnung der Klassenzimmertüren“

Unser Ansatz der Unterrichtsdiagnostik zielt eindeutig auf *Reflexion* und Austausch im kollegialen Umfeld ab. Das zugrunde liegende Leitbild ist das des „reflective practioner“ (Schön, 1983), der seinen Unterricht datenbasiert erforscht. Im Unterschied zu Unterrichtsbeobachtungen im Rahmen der Externen Evaluation geht es hier *nicht* darum, den Unterricht so objektiv wie möglich zu beschreiben, sondern darum, Gesprächsanlässe für eine Verständigung über Unterricht zu schaffen. Für eine Benotung oder Bewertung des Unterrichts im Rahmen von Personalbeurteilungen oder Lehrproben ist EMU nicht geeignet.

Hier finden Sie mehr zum Potenzial der Unterrichtsdiagnostik für Schule und Lehrerfortbildung, für Studienseminare sowie für die universitäre Lehrerausbildung.

4.3 Welchen wissenschaftlichen Hintergrund hat EMU?

Mit Diagnostik ist ein höherer Anspruch verbunden als mit einer intuitiven Eindrucksbildung: Grundlage der Beobachtung sind wissenschaftlich fundierte, d.h. empirisch gut untersuchte Merkmale der Unterrichtsqualität. Günstige Ausprägungen dieser Merkmale sind nachweislich lernwirksam (Hattie, 2012). Außerdem müssen diagnostische Instrumente – im Gegensatz zu ad-hoc entwickelten Verfahren – bestimmte methodische Mindeststandards erfüllen und in der Praxis erprobt worden sein.

Gegenstand von EMU sind nicht Methoden, sondern die „Prinzipien des effektiven Lehrens und Lernens“ (Hattie, 2009), und zwar vier zentrale fachübergreifende *Prozessmerkmale* der Unterrichtsqualität: (1) Effiziente Klassenführung, (2) Lernförderliches Klima und Motivierung, (3) Klarheit und Strukturiertheit und (4) Kognitive Aktivierung. Hinzu kommt (5) ein *Bilanzbereich*, d.h. eine Einschätzung der Stunde in emotionaler, (Wohlfühlen), motivationaler (Interessantheit) und kognitiver Hinsicht (Lernertrag, Passung).

Hier gibt es Informationen zum wissenschaftlichen Hintergrund der Unterrichtsdiagnostik und zur Qualität des Unterrichts. Grundlage ist das Lehrbuch *Unterrichtsqualität und Lehrerprofessionalität – Diagnose, Evaluation und Verbesserung des Unterrichts*) (2014) von A. Helmke.

4.4 Was heißt „Abgleich von Perspektiven“?

Die folgende Übersicht veranschaulicht das Prinzip der Erfassung des Unterrichts durch äquivalente Angaben aus unterschiedlichen Perspektiven (unterrichtende Lehrkraft, Schüler/in, Kollege/in). Exemplarisch für eine weibliche Lehrkraft wird für jeden Bereich jeweils eine Frage aus allen Perspektiven dargestellt: (1) Klassenführung, (2) Lernförderliches Klima, (3) Klarheit/Strukturiertheit, (4) Aktivierung und (5) Bilanz.

Schülerfragebogen	Lehrerfragebogen	Kollegenfragebogen
Ich konnte in dieser Unterrichtsstunde ungestört arbeiten.	Die Schüler/innen konnten ungestört arbeiten.	Die Schüler/innen konnten ungestört arbeiten.
Wenn die Lehrerin in dieser Unterrichtsstunde eine Frage gestellt hat, hatte ich ausreichend Zeit zum Nachdenken.	Wenn ich eine Frage gestellt habe, hatten die Schüler/innen ausreichend Zeit zum Nachdenken.	Wenn die Kollegin eine Frage gestellt hat, hatten die Schüler/innen ausreichend Zeit zum Nachdenken.
Mir ist klar, was ich in dieser Stunde lernen sollte.	Den Schüler/innen war klar, was sie in dieser Stunde lernen sollten.	Den Schüler/innen war klar, was sie in dieser Stunde lernen sollten.
Ich war die ganze Stunde über aktiv bei der Sache.	Die Schüler/innen waren die ganze Stunde über aktiv bei der Sache.	Die Schüler/innen waren die ganze Stunde über aktiv bei der Sache.
Ich habe in dieser Unterrichtsstunde etwas dazugelernt.	Die Schüler/innen haben in dieser Stunde etwas dazugelernt.	Die Schüler/innen haben in dieser Stunde etwas dazugelernt

Der Abgleich der eigenen Sichtweise mit den Schülerangaben ist ein Schritt, um das Lernen sichtbar zu machen, d.h. das Lernen mit den Augen der Schüler zu sehen: *„Lernen muss von den Lehrpersonen aus der Perspektive der Lernenden betrachtet werden, damit sie besser verstehen, wie das Lernen aus der Sicht der Lernenden aussieht und wie es sich für sie anfühlt“* (Hattie, 2013, S. 139). Die Fragen sollen ein Katalysator dafür sein, *„dass Lehrpersonen nach widerlegbaren empirischen Belegen zur Effektivität ihres Unterrichtens suchen, dass sie nach Irrtümern in ihrem Wissen und Ihren Vorstellungen suchen, ... dass sie fragen, ob es genug Herausforderungen und Engagement beim Lernen gibt“* (Hattie, 2013, S. 298).

Der EMU-Schülerfragebogen weist zwei Besonderheiten auf, verglichen mit anderen Schülerfragebögen zum Unterricht:

- ***Anschlussfähigkeit***: Gegenstand ist nicht der Unterricht im Allgemeinen bzw. über einen längeren Zeitraum, sondern eine *konkrete Unterrichtsstunde*. Dies ermöglicht den Abgleich der Schülersicht mit denjenigen des unterrichtenden und hospitierenden Kollegen. Um den Einsatz des Schülerfeedbacks auch für längere Referenzzeiträume (z.B. Unterrichtseinheit, Lernsituation) zu ermöglichen, gibt es zusätzlich eine WORD-Version der Fragebögen, die entsprechend adaptiert werden kann.
- ***Subjektivität***: Die Items verwenden überwiegend die „Ich“-Form statt „wir“ oder „uns“. Schüler müssen sich also nicht in die Perspektive ihrer Mitschüler versetzen und eine „Durchschnittsbildung“ vornehmen, sondern beurteilen ihr eigenes subjektives Erleben. Da ein- und dasselbe Unterrichtsangebot je nach individuellen Lernvoraussetzungen oft ganz unterschiedlich wahrgenommen, interpretiert und genutzt wird, kann die Sichtung solcher Ergebnisse für Heterogenität sensibilisieren.

Der Abgleich schafft Anlässe, um gemeinsam über Verlauf und Ertrag der Unterrichtsstunde, über Konsens und Dissens bei der Beurteilung nachzudenken: *„Hier diskutieren, bewerten und planen sie ihren Unterricht im Licht der Feedback-Evidenz ... Dies ist nicht (nur) kritische Reflexion, sondern kritische Reflexion im Licht der Evidenz, also im Licht empirischer Belege zu ihrem Unterricht“* (Hattie, 2013, S. 281).

Das Instrument ist modular aufgebaut. Man kann z.B.

→ das Instrument zunächst einmal nur für sich selbst, als eine Art „Logbuch“ verwenden
→ einen *Überblick* über alle fünf Bereiche gewinnen, aber auch nur *einen* Bereich auswählen
→ das Instrument nur *punktuell* („Momentaufnahme“) oder zur Erfassung von *Veränderungen* als Ergebnis der Unterrichtsentwicklung mehrfach einsetzen
→ die eigene Sicht mit *einer* anderen Perspektive statt mit beiden (Kollege, Klasse) abgleichen

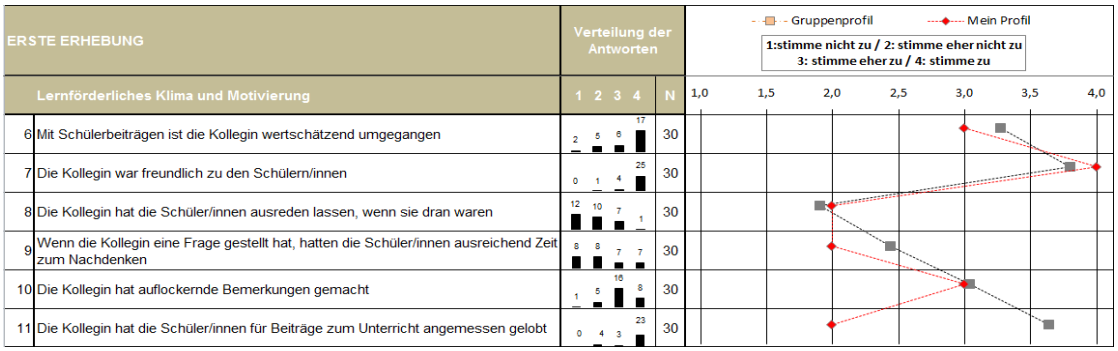
Das o.g. Fundamentum kann um Zusatzbereiche („Additum“) ergänzt werden, z.B.:

→ Kognitive Aktivierung, Umgang mit Vielfalt, effiziente Gruppenarbeit, Lehrersprache, Feedbackverhalten und Orientierung an den Bildungsstandards
→ Instrumente zum Unterricht der Externen Evaluation / Schulinspektion des Bundeslandes
→ selbstentwickelte Items zu Bereichen, die der Schule wichtig sind
→ Beobachtungsaufträge, z.B. zum mündlichen Sprachverhalten von Schülern

Hier gelangen Sie zu den Fragebögen. Es gibt Versionen für eine männliche vs. weibliche Lehrperson, um politisch korrekte, aber unelegante Formulierungen wie „Der Lehrer / die Lehrerin“ zu vermeiden.

4.5 Was leistet das Auswertungsprogramm?

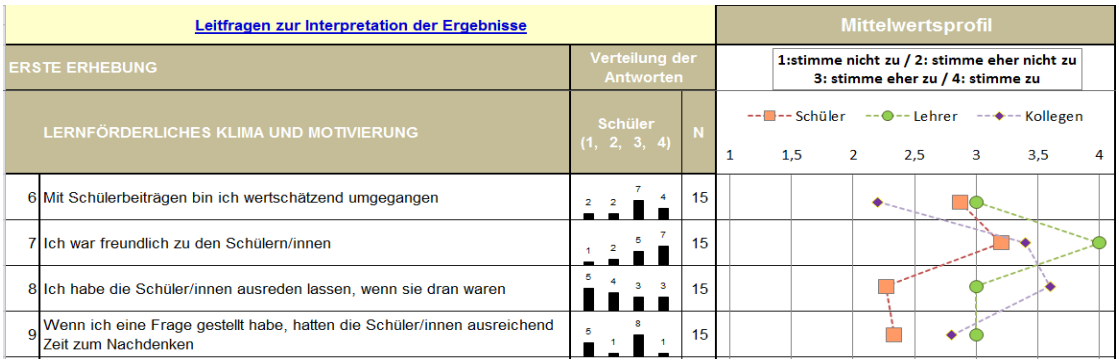
Geht es lediglich um den Abgleich zwischen unterrichtender und hospitierender Lehrperson, braucht man keine Software, sondern kann einfach die beiden Bögen nebeneinander halten und vergleichen. Das Potenzial der Software liegt in der Visualisierung von Ergebnissen, wenn zu ein und demselben Unterricht *viele* Urteile vorliegen, wie beim Schülerfeedback oder der videobasierten Einschätzung einer Unterrichtssequenz durch eine Gruppe (siehe die untenstehende Abbildung).



Nach der Eingabe der Daten visualisiert das Programm auf Knopfdruck *erstens* die Verteilung der Antwortkategorien (von 1 = *stimme nicht zu* bis 4 = *stimme zu*) in Form von Stabdiagrammen, um so Konsens und Dissens zu veranschaulichen. *Zweitens* stellt das Programm das individuelle Urteilsprofil dem Durchschnittsprofil der Gesamtgruppe gegenüber: Wo bin ich mit meinem Urteil im *mainstream*, wo weiche ich vom Durchschnitt ab? Im obigen Fall hat die Kollegin („Mein Profil") möglicherweise ein spezifisches Konzept von „angemessenem Lob".

Hier finden Sie Anregungen für die Unterrichtsanalyse im Team.

Beim Datenabgleich aus drei Perspektiven (Triangulation) könnte eine Visualisierung so aussehen:



Beim Schülerfeedback ist die Antwortverteilung *innerhalb* der Klasse oft interessanter als der Klassenmittelwert, siehe die Stabdiagramme auf der linken Seite der Abbildung. Diese Verteilung ist ein Ausdruck für Homogenität oder Heterogenität innerhalb der Klasse, denn je nach individuellen Lernvoraussetzungen wird ein und dasselbe Unterrichtsangebot unterschiedlich wahrgenommen, interpretiert und genutzt. Im obigen Beispiel gibt es einen deutlichen Konsens bei Items 6 und 7, die sich auf ein durch Freundlichkeit und Wertschätzung charakterisiertes Klima beziehen. Für die Items 8 und 9 dagegen zeigt sich eine erhebliche klasseninterne Streuung: Die Wartezeit nach Fragen wird von der Mehrheit der Schüler/innen für ausreichend gehalten, ein beachtlicher Teil der Schüler/innen hätte jedoch mehr Zeit zum Nachdenken gebraucht.

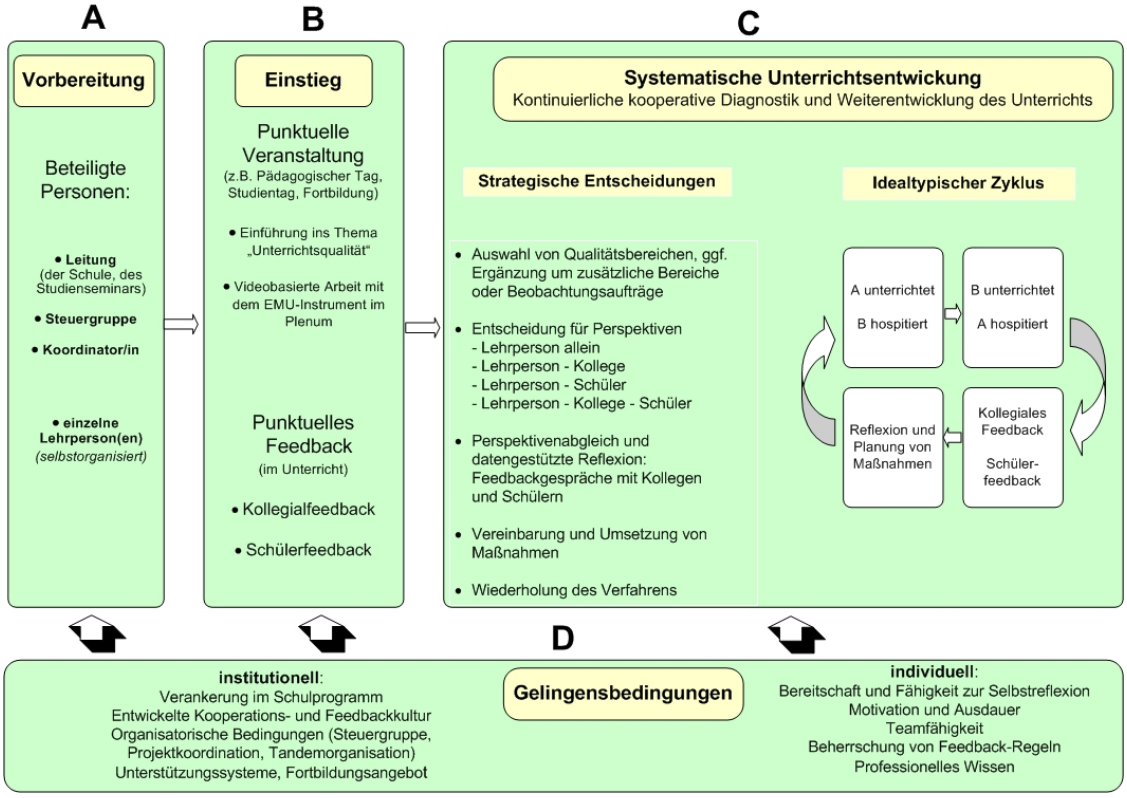
Wie aufwändig ist die Dateneingabe? Die Erfahrung zeigt, dass man pro Schüler höchstens 1 Minute für die Dateneingabe benötigt. Bei Daten von 30 Schülern dauert es also höchstens eine halbe Stunde. In der Praxis delegieren die meisten Lehrpersonen die Dateneingabe an Schüler in ihrer Klasse, die solche Arbeit gerne und kompetent erledigen. Zur Unterstützung der datenbasierten Reflexion über Unterricht haben wir beispielbasierte Leitfragen entwickelt.

EMU ist ein Offline-Verfahren. Das heißt: Das gesamte Material befindet sich im Netz (und kann heruntergeladen werden), aber die konkreten Daten werden „traditionell" durch die schriftliche Bearbeitung von Fragebögen und Checklisten erzeugt und anschließend in die zur Verfügung gestellte Dateneingabemaske eingetragen. Schulen, die bereits über Online-Werkzeuge verfügen (wie z.B. UniPark, LimeSurvey, GrafStat), können die EMU-Fragebögen in ihre Verfahren einpflegen.

Hier geht es zu den Auswertungsprogrammen und den Manualen.

4.6 Welche Szenarien und Veranstaltungsformate haben sich in der Praxis bewährt?

Die folgende Abbildung veranschaulicht die Gesamtarchitektur und einzelne Szenarien der Unterrichtsdiagnostik. Dabei werden drei Phasen (A, B, C) unterschieden:



A) **Vorbereitung**: Die Initiierung der Unterrichtsdiagnostik als Katalysator der Unterrichtsentwicklung erfordert ein systematisches und koordiniertes Vorgehen und ist eine Führungsaufgabe für die Schulleitung. Die Vorbereitung erfolgt zweckmäßigerweise mit kollegialer Unterstützung, z.B. einer Steuergruppe.

B) Der **Einstieg** in die Unterrichtsdiagnostik (B) kann auf unterschiedliche Weise erfolgen:

- Für die ***Schule*** empfiehlt sich zu Beginn eine zentrale Veranstaltung (z.B. im Rahmen eines Pädagogischen Tages oder einer SchiLF), die wie folgt verlaufen kann: in die Thematik einführen, eine videografierte Unterrichtssequenz zeigen, diese mit dem EMU-Bogen beurteilen, die Daten eingeben und datenbasiert über Dissens und Konsens diskutieren. Als Hilfestellung für eine solche Veranstaltung steht eine PowerPoint-Präsentation zur Verfügung. Außerdem werden Hinweise auf erhältliche Unterrichtsvideos gegeben. Denkbar ist jedoch auch ein Start in Gestalt einer punktuellen Nutzung der EMU-Instrumente durch einzelne Lehrkräfte / Tandems im Rahmen einer kollegialen Hospitation (Individualfeedback) oder durch Schülerfeedback.
- ***Studienseminaren*** bietet die Arbeit mit EMU vielfältige Lernchancen. Bewährt hat sich eine halbtägige Fortbildungsveranstaltung, günstigenfalls gefolgt von einem längerfristigen Projekt.
- Schließlich ist die Unterrichtsdiagnostik ein möglicher Gegenstand ***universitärer*** Lehrerausbildung.

C) **Systematische Unterrichtsentwicklung**: Das hauptsächliche Potenzial von EMU liegt in einem längerfristigen Programm der Diagnostik und Reflexion des Unterrichts, gekoppelt mit systematischer Unterrichtsentwicklung. Wo es bereits eine entwickelte Kultur der Kooperation gibt, können die Phasen A und B auch entfallen.

Die Abbildung zeigt (auf der rechten Seite) einen idealtypischen Verlauf der Unterrichtsdiagnostik durch ein Tandem, bei dem systematisch die Rollen gewechselt werden. Das Vorgehen sollte den bewährten Dreischritt „Bestandsaufnahme – Intervention – Evaluation" zugrunde legen, d.h. die erste Erhebung versteht sich als Screening von Stärken und Schwächen im Unterricht und bildet die Grundlage für Planung von Maßnahmen der Professionalisierung (z.B. vertiefende Information) und der Weiterentwicklung des Unterrichts. Ob diese erfolgreich war, kann durch eine Wiederholung der Erhebung zu einem späteren Zeitpunkt festgestellt werden.

Wechselseitige Unterrichtsbesuche sind das Herzstück von EMU. Neben ihrer diagnostischen Funktion für die Weiterentwicklung des Unterrichts können so Impulse für die **Schulentwicklung** erfolgen: weg von der Einzelkämpfermentalität, von der noch immer vorherrschenden Vorstellung des Unterrichts als Privatangelegenheit hin zu einer professionellen Lerngemeinschaft.

4.7 Wovon hängt das Gelingen ab?

Der Erfolg des Unternehmens hängt von wichtigen institutionellen und individuellen Bedingungen ab:

- Auf ***institutioneller*** Seite ist es günstig, wenn die Unterrichtsdiagnostik im Schulprogramm verankert ist und bereits eine innerschulische Feedbackkultur besteht. Wichtig ist dabei sind unterrichtswirksame Schulleitungen *„die für herausfordernde Ziele eintreten und dann ein sicheres Umfeld für Lehrpersonen schaffen, in dem Kritik, Fragen und die Unterstützung anderer Lehrpersonen möglich sind. So können diejenigen Ziele gemeinsam erreicht werden, die den größten Effekt auf Schüler-Outcomes haben“* (Hattie, 2013, S. 99). Als wichtig erwies sich, dass ausreichende Ressourcen zur Verfügung stehen, insbesondere Zeit für Hospitation und anschließende Reflexion.
- Auf ***individueller*** Ebene erfordert die erfolgreiche Durchführung der Diagnostik die Fähigkeit und Bereitschaft zur Selbstreflexion, gekoppelt mit der Fähigkeit, im Team zu arbeiten. Um die zurückgemeldeten Unterrichtsbeurteilungen verstehen zu können, ist ein Mindestmaß an Vertrautheit mit graphischen und tabellarischen Darstellungen empirischer Ergebnisse erforderlich. Weiterhin muss es für eine Lehrperson lohnenswert sein, sich an der Unterrichtsdiagnostik zu beteiligen, d.h. der erhoffte Nutzen muss größer sein als befürchtete Kosten (Zeitverlust, Verunsicherung).
- Für die Arbeit der ***Tandems*** ist die Kenntnis wichtiger Regeln des Gebens und Nehmens von Feedback unabdingbar.

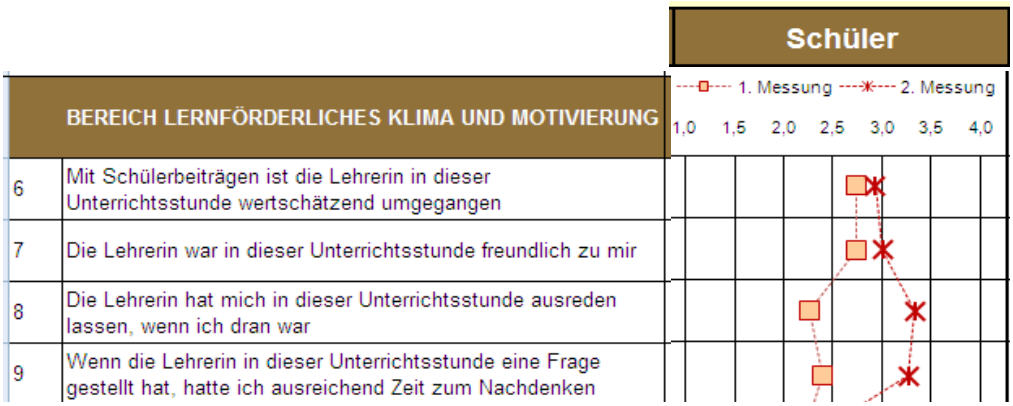
Die Erfahrungen in der Praxis zeigen, dass es bei der Unterrichtsdiagnostik auch „Stolpersteine“ gibt. Sie lassen sich zwar nicht immer vollständig ausräumen, man kann sie aber zumindest entschärfen – wenn man sie kennt und sich rechtzeitig darauf einstellt. Hier finden Sie Hinweise auf häufig gehörte Einwände sowie Möglichkeiten, damit umzugehen.

4.8 Unterrichtsdiagnostik – und was dann?

Wer kennt nicht Sprüche wie „Vom Wiegen wird die Sau nicht fetter“ oder „Entwickeln statt vermessen!“? Das Fatale an diesen populistischen Floskeln ist, dass sie einen wahren Kern enthalten: Diagnostik ist kein Selbstzweck, sondern ihr müssen zielgerichtete Maßnahmen nachfolgen („Von Daten zu Taten“). Beides ist nötig: eine solide Standortbestimmung und daraus abgeleitete Konsequenzen. Hierfür steht EMU ein Formblatt für die Protokollierung und Vereinbarung von Maßnahmen zur Verfügung.

Ebenso wie für qualitativ hochwertigen Unterricht gibt es auch für die Unterrichtsentwicklung keinen Königsweg. Je nach Sachlage und vorhandenen Ressourcen kommt die gesamte Bandbreite von Maßnahmen der Unterrichtsentwicklung in Betracht: von überregionaler bis hin zu schulinterner Fortbildung, von Lehrerverhaltenstrainings bis hin zu Methoden des Coaching, vom Training des Umgangs mit Disziplinproblemen bis zur Förderung von Methodenkompetenzen, vom Microteaching bis hin zum Lernen aus Videos. Wichtig: Ein kriteriengeleiteter und datenbasierter, gut vor- und nachbereiteter Austausch über beobachteten Unterricht ist selbst eine der effektivsten Formen der Lehrerfortbildung überhaupt!

Die EMU-Software ermöglicht die Analyse einer Messwiederholung. So könnte man sich als Ergebnis einer Bestandsaufnahme (Messung 1) vornehmen, seinen Unterricht gezielt zu verändern. Messung 2 (ausreichender zeitlicher Abstand, gleiches Fach, ähnlicher Stundentypus) kann dann Veränderungen visualisieren, siehe die folgende Abbildung. Dort war eine Lehrperson mit dem Ergebnis der ersten Erhebung unzufrieden und nahm sich vor, geduldiger zu sein – mit Erfolg, wie der Unterschied zwischen den beiden Profilen zeigt.



Hier finden Sie Leitfragen und Beispiele zur Interpretation von Veränderungen.

Die Feststellung quantitativ darstellbarer Veränderungen bei bestimmten Unterrichtsmerkmalen ist aber nicht das einzige und vielleicht nicht einmal der wichtigste Ziel der Unterrichtsdiagnostik. Die Konfrontation der Selbsteinschätzungen mit anderen Sichtweisen und das Gespräch mit einem sachkundigen und kritischen, aber wohlwollenden Partner oder der Klasse ist eine ausgezeichnete Lerngelegenheit, um sich eigener Sichtweisen, Erklärungen und Verhaltensmuster klar zu werden. Subjektive Theorien des Lehrens und Lernens steuern zwar das Handeln, bleiben aber oft unterhalb der Schwelle des bewussten Nachdenkens. Der Abgleich hat das Potenzial, solche intuitiven Konzepte der bewussten Kontrolle zugänglich zu machen und implizite Theorien explizit zu machen. Dies ist eine günstige Voraussetzung für die Ingangsetzung von Veränderungsprozessen. EMU hat in vielen Lehrerzimmern bewirkt, dass dort seit langem erstmalig intensiv und engagiert über pädagogische und didaktische Fragen des Unterrichts gesprochen wurde!

4.9 Wie kann das Kollegium zum Mitmachen motiviert werden?

Es wäre unrealistisch zu erwarten, dass sich ein Kollegium spontan und enthusiastisch mit der Unterrichtsdiagnostik beschäftigt. Es kommt also entscheidend darauf an, ob die Leitung das Kollegium vom Sinn einer Teilnahme überzeugen kann, denn Unterrichtsdiagnostik funktioniert nicht per Anordnung, sondern erfordert eine tragfähige Basis im Kollegium.

Im Folgenden sind einige Argumente für eine Teilnahme stichwortartig aufgeführt:

- Appell an die Professionalität der Lehrerkolleginnen und -kollegen
- Erleben der Wirksamkeit des eigenen Unterrichts als Schutz vor Erschöpfung („burn-out“)
- Die Schulleitung nimmt selbst aktiv teil und ist damit Vorbild („Lernen am Modell“)
- Kollegiales Feedback als Schritt zur Entwicklung einer innerschulischen Kooperationskultur
- Hinweis auf Vorgaben und Empfehlungen in Schulgesetzen und Orientierungsrahmen
- Schülerfeedback, um unseriösen Praktiken („spick-mich.de“) den Wind aus den Segeln zu nehmen
- Wertschätzung innovativer Vorhaben durch Schulleitung, Schulaufsicht und Schulträger
- Unterrichtsdiagnostik als ein Schritt in Richtung Exzellenz (Zertifizierung, Gütesiegel)
- hohe Wertschätzung von Initiativen der Unterrichtsdiagnostik seitens der Eltern

4.10 EMUplus: Unterrichtsdiagnostik und Lehrergesundheit

Guter Unterricht, bei dem die Schüler/innen viel lernen, der sie unter Berücksichtigung ihrer Verschiedenheit individuell fördert und in einem lernförderlichen Klima stattfindet, steigert die Zufriedenheit und das Erleben der Wirksamkeit der Lehrpersonen und ist somit zugleich ein wirksamer Schutz vor Erschöpfung. Guter Unterricht *allein* ist allerdings keine Garantie für den Erhalt der Lehrergesundheit, denn Überengagement, unrealistisch hohe Erwartungen, schwierige Schüler/innen, Lärm, mangelnde Unterstützung im Kollegium sind gravierende berufliche Belastungsfaktoren.

Wir haben deshalb – in Kooperation mit dem Kultusministerium des Landes Baden-Württemberg – ein Zusatzmodul entwickelt, das anders als die bisherigen Module *qualitativen* Charakter hat, also nicht zu einem datenbasierten Ergebnisabgleich mit anschließender Visualisierung führt, sondern Grundlage für ein kollegiales Gespräch ist, siehe EMUplus.

4.11 Literatur (Stand: 15.06.2014)

im Druck

Helmke, A. (im Druck). Unterrichtsdiagnostik als Ausgangspunkt von Unterrichtsentwicklung. In H.G. Rolff (Hrsg.), *Handbuch Unterrichtsentwicklung*. Beltz.

Helmke, A., Helmke, T., Lenske, G., Pham, G., Praetorius, A.-K., Schrader, F.-W. & Ade-Thurow, M. (im Druck). Unterrichtsdiagnostik mit EMU. In M. Ade-Thurow, W. Bos et al. (Hrsg.*). Aus- und Fortbildung der Lehrkräfte in Hinblick auf Verbesserung der Diagnosefähigkeit, Umgang mit Heterogenität und individuelle Förderung*. Münster: Waxmann.

Helmke, A. & Lenske, G. (im Druck). Selbstreflexion und kollegialer Austausch als Elemente der Lehrerprofessionalität. *Grundschulmagazin*.

Lenske, G. & Helmke, A. (in Druck). Child respondents – do they really answer to what scientific questionnaires ask for? In W. Schnotz, A. Kauertz, H. Ludwig, A. Müller, J. Pretsch (2014): *Multidisciplinary Research on Teaching and Learning*. Basingstoke: Palgrave Macmillan.

2014

Ade-Thurow, M. (2014). Erfahrungen einer Schule mit evidenzbasierten Methoden der Unterrichtsdiagnostik und -entwicklung (EMU). In: C. Fischer (Hrsg.), *Damit Unterricht gelingt. Von der Qualitätsanalyse zur Qualitätsentwicklung*. Münster: Waxmann.

Ade-Thurow, M (2014a). Unterricht mit den Augen der Schüler sehen. Erfahrungen mit dem Instrument EMU in der Sekundarstufe I. *Pädagogik, 4*/14, 24-29.

Helmke, A. & Helmke, T. (2014). Unterrichtsanalyse mit EMU (Evidenzbasierte Methoden der Unterrichtsentwicklung). *Journal für Schulentwicklung*, 18, 1/2014, 55-57.

Helmke, A. & Helmke, T. (2014). EMU – Systematische Entwicklung der Unterrichtsqualität. *Lernende Schule*, 66, 34-36.

Helmke, A. & Helmke, T.(2014). Unterrichtsdiagnostik mit EMU. *Schulblatt des Kantons Thurgau*, 56, April 2014, 16-19.

Lenske, G. (2014). *Schülerfeedback in der Grundschule: Studien zur Validität*. Unveröffentlichte Dissertation. Universität Koblenz-Landau.

Praetorius, A.-K. (2014). *Messung von Unterrichtsqualität durch Ratings.* Münster: Waxmann.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K. & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2-12. doi: 10.1016/j.learninstruc.2013.12.002.

2013

Helmke, A. & Lenske, L. (2013). Unterrichtsdiagnostik als Voraussetzung für Unterrichtsentwicklung. *Beiträge zur Lehrerbildung,* Beiträge zur Lehrerbildung, 31 (2), 2013, 214-233.

Helmke, A., Helmke, T., Lenske, G., Pham, G., Praetorius, A.-K., Schrader, F.-W. & Ade-Thurow, M. (2013). EMU – Unterrichtsdiagnostik. Studienbrief Version 4.0. Kultusministerkonferenz: Projekt EMU (Evidenzbasierte Methoden der Unterrichtsdiagnostik). Landau: Universität Koblenz-Landau, Campus Landau.

Helmke, A., Helmke, T. & Pham, G. (2013). Unterrichtsdiagnostik als künftige Aufgabe für Lehrerinnen und Lehrer. In S. Lin-Klitzing, D. Di Fuccia & G. Müller-Frerich (Hrsg.), *Zur Vermessung von Schule. Empirische Bildungsforschung und Schulpraxis (S. 153-165).* Reihe Gymnasium – Bildung – Gesellschaft. Bad Heilbrunn: Klinkhardt.

Helmke, A. & Pham, G. (2013). Unterrichtsdiagnostik. In M. A. Wirtz, H. O. Häcker & K.-H. Stapf (Hrsg.), *Dorsch – Lexikon der Psychologie (S. 1602)*. Bern: Huber.

Helmke, A., Helmke, T. & Pham, G. (2013). Unterrichtsqualität und Unterrichtsdiagnostik. In A. Besand (Hrsg.), *Lehrer- und Schülerforschung in der politischen Bildung*. Schwalbach: Wochenschauverlag.

Pham, G. & Helmke, A. (2013). Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung (EMU). In M. A. Wirtz, H. O. Häcker & K.-H. Stapf (Hrsg.), *Dorsch – Lexikon der Psychologie (S. 507)*. Bern: Huber.

Praetorius, A.-K. (2013). Einschätzung von Unterrichtsqualität durch externe Beobachterinnen und Beobachter – Eine kritische Betrachtung der aktuellen Vorgehensweise in der Schulpraxis. *Beiträge zur Lehrerbildung, 31*, 174-185.

Schrader, F.-W. & Helmke, A. (2013). EMU – Selbst- und Fremdbeurteilung von Unterricht. *Pädagogische Führung, Thema „Beurteilung". 24* (3), 88-91.

2012

Helmke, A. (2012). Unterricht diagnostizieren und evaluieren. Voraussetzungen für die Verbesserung der Unterrichtsqualität. *SchulVerwaltung spezial, 14* (4), 16-19.

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* (4. überarbeitete Aufl., Schule weiterentwickeln – Unterricht verbessern. Orientierungsband). Seelze: Klett-Kallmeyer.

Helmke, A., Helmke, T. & Pham, G. (2012). Unterrichtsfeedback. In T. Riecke-Baulecke (Hrsg.), *Schulmanagement-Handbuch 144: Interne Evaluation* (Vol. 31, S. 45-64). München: Oldenbourg.

Helmke, A., Helmke, T. & Pham, G. (2012). Von der externen zur internen Evaluation des Unterrichts. *Hamburg macht Schule, 3*, 31-32.

Helmke, A., Helmke, T. & Schrader, F.-W. (2012). EMU. Von der Unterrichtsdiagnostik zur Unterrichtsentwicklung. *Friedrich Jahresheft XXX 2012 „Schule vermessen"*, 122-124.

Helmke, A., Helmke, T. & Schrader, F.-W. (2012). Unterrichtsdiagnostik mit EMU - Unterricht aus mehreren Perspektiven betrachten und diskutieren. PDF Download Unterrichtsqualität Sekundarstufe In M. Bonsen, W. Homeier & K. Tschekan (Hrsg.), *Unterrichtsqualität sichern – Sekundarstufe. 17. Ergänzung*. Stuttgart: Raabe Verlag.

Helmke, A., Helmke, T., Lenske, G., Pham, G., Praetorius, A.-K., Schrader, F.-W. & Ade-Thurow, M. (2012). Unterrichtsdiagnostik – Voraussetzung für die Verbesserung der Unterrichtsqualität. In S. G. Huber (Hrsg.), *Jahrbuch Schulleitung 2012. Befunde und Impulse zu den Handlungsfeldern des Schulmanagements* (S. 133-144). Köln: Carl Link.

Helmke, A., Piskol, K., Pikowsky, B. & Wagner, W. (2012). Schüler als Experten von Unterricht. Unterrichtsqualität aus Schülerperspektive. In Pädagogisches Landesinstitut Rheinland-Pfalz (Hrsg.), *Mit Heterogenität umgehen – Differenziert unterrichten in der Sekundarstufe. Sammelband* (S. 10-17). Velber: Friedrich Verlag. (Nachdruck aus Lernende Schule, 46-47, 98-105).

Helmke, A., Schrader, F.-W. & Helmke, T. (2012). EMU: Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung. Unterrichtdiagnostik – Ein Weg, um Unterrichten sichtbar zu machen. *Schulverwaltung Baden-Württemberg, 21* (7/8), 167-170.

Helmke, A., Schrader, F.-W. & Helmke, T. (2012). EMU: Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung. Unterrichtsdiagnostik – Ein Weg, um Unterrichten sichtbar zu machen. *Schulverwaltung Bayern, 35* (6), 180-183.

Helmke, A., Schrader, F.-W. & Helmke, T. (2012). EMU: Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung. Unterrichtsdiagnostik – Ein Weg, um Unterrichten sichtbar zu machen. In Staatsinstitut für Schulqualität und Bildungsforschung (ISB) (Hrsg.), *Einblicke – Ausblicke. Jahrbuch 2011* (S. 125-135). München: Staatsinstitut für Schulqualität und Bildungsforschung (ISB).

Helmke, A., Schrader, F.-W., Helmke, T. & Pham, G. (2012). EMU: Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung. Unterrichtsdiagnostik – Ein Weg, um Unterrichten sichtbar zu machen. *Schulverwaltung Nordrhein-Westfalen, 23* (12), 325-328.

Pham, G., Koch, T., Helmke, A., Schrader, F.-W., Helmke, T. & Eid, M. (2012). Do teachers know how their teaching is perceived by their pupils? *Procedia – Social and Behavioral Sciences Journal, 46*, 3368 – 3374.

Praetorius, A.-K., Lenske, G. & Helmke, A. (2012). Observer Ratings of Instructional Quality: Do They Fulfill What They Promise?, *Learning and Instruction,* doi:10.1016/j.learninstruc.2012.03.002.

2011

Ade-Thurow, M. (2011). *Empirische Untersuchungen zur Unterrichtsdiagnostik an Schulen.* unveröffentlichte Masterthesis im Masterstudiengang Schulentwicklung, Pädagogische Hochschule in Weingarten, Weingarten.

Helmke, A. (2011). Besser unterrichten mit EMU. Wissenschaftler und Praktiker entwickeln neue Methode zur Unterrichtsdiagnostik. *Schule im Blickpunkt. Informationen des Landeselternbeirats Baden-Württemberg, 44* (4), 17-19.

Helmke, A. & Helmke, T. (2011). Diagnostische Kompetenzen. Unterrichtsanalyse mit EMU. *SCHULE NRW 06/11. Amtsblatt des Ministeriums für Schule und Weiterbildung*, 288-290.

Helmke, A. & Helmke, T. (2011). Unterrichtsdiagnostik. Eine Frage der Perspektive. *INFO Informationsschrift für Kindergarten und Schule in Südtirol, Dezember 2011*, 12-13.

Helmke, A., Helmke, T., Lenske, G., Pham, G., Praetorius, A.-K., Schrader, F.-W. & Ade-Thurow, M. (2011). EMU – Unterrichtsdiagnostik. Studienbrief Version 2.03. Kultusministerkonferenz: Projekt EMU (Evidenzbasierte Methoden der Unterrichtsdiagnostik). Landau: Universität Koblenz-Landau, Campus Landau.

Helmke, A., Helmke, T., Lenske, G., Pham, G., Praetorius, A.-K., Schrader, F.-W. & Ade-Thurow, M. (2011). Unterrichtsdiagnostik – Voraussetzung für die Verbesserung der Unterrichtsqualität. In A. Bartz, M. Dammann, S. Huber, C. Kloft & M. Schreiner (Hrsg.), *PraxisWissen SchulLeitung, AL 28, 2011 (Kap. 30.71)*. Köln: Wolters Kluwer.

Helmke, A., Helmke, T. & Schrader, F.-W. (2011). Unterrichtsdiagnostik. Unterstützung für die Schulpraxis. *schulmanagement, 4-2011* (4), 15-17.

Helmke, A. & Schrader, F.-W. (2011). Unterrichtsqualität: Von der Unterrichtsdiagnostik zur Unterrichtsentwicklung. *Schule heute, 51* (6), 8-9.

Helmke, A. & Schrader, F.-W. (2011). Unterrichtsqualität: Von der Unterrichtsdiagnostik zur Unterrichtsentwicklung. *VBE zeitnah, 4* (7-9), 13-14.

Schrader, F.-W. (2011). Lehrer als Diagnostiker. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 683-698). Münster: Waxmann.

2010

Helmke, A., Helmke, T., Lenske, L., Pham, G. H., Praetorius, A.-K., Schrader, F.-W. & Ade-Thurow, M. (2010). *Studienbrief Unterrichtsdiagnostik. EMU. Evidenzbasierte Methoden der Unterrichtsdiagnostik*, [pdf + program]. Universität Koblenz-Landau, Campus Landau. Information: www.unterrichtsdiagnostik.info [2012, 29.03.].

2009

Helmke, A., Piskol, K., Pikowsky, B. & Wagner, W. (2009). Schüler als Experten von Unterricht. Unterrichtsqualität aus Schülerperspektive. *Lernende Schule* (46-47), 98-105

2008

Helmke, A. (2008). Wie können Lehrkräfte ihren Unterricht reflektieren und bewerten? *SCHULE NRW 09/08. Amtsblatt des Ministeriums für Schule und Weiterbildung* (60. Jahrgang Nr. 9), 430-434.