Authentizität bei der Beurteilung von Fachleistungen und Lernkompetenzen

Kurt Reusser und Rita Stebler

Die Ausläufer der Amerikanischen Assessment Debatte dringen allmählich in den deutschen Sprachraum vor. In ihrem Gefolge erscheinen Testinstrumente, die entwickelt wurden, um Fachleistungen und Lernkompetenzen in komplexen Anwendungssituationen zu messen. Zwei Beispiele, Experimentieraufgaben und Testportfolios, werden in diesem Beitrag vorgestellt. Die Arrangements entsprechen einem erweiterten Verständnis von Lernen und Lehren und können in Kombination mit herkömmlichen Leistungstests zu höherer Authentizität beim Prüfen führen.

Im Bereich der Leistungsmessung und -beurteilung macht ein Schlagwort die Runde. Es heisst *Authentizität* (Wiggins, 1989a) und steht für eine zentrale Forderung in der Amerikanischen Assessment Debatte (Terwilliger, 1997). In diesem Beitrag erörtern wir diese Forderung aus lernpsychologischer und didaktischer Perspektive, um dann auf zwei Formen der Leistungsmessung einzugehen, die oft mit dem Gütesiegel 'authentisch' versehen werden. Wir wählen dabei ein praxisnahes Vorgehen, das die eine oder andere Lehrperson zu Umsetzungsversuchen anregen könnte. Schliesslich werden einige Vorzüge und Fallstricke dieser 'authentischen' Testverfahren erörtert.

1 Authentizität als Gütekriterium der Leistungsmessung und -beurteilung

In den USA nehmen jedes Jahr Tausende Schülerinnen und Schüler an obligatorischen Leistungsvergleichen teil. Aus test- und finanzökonomischen Gründen werden dabei hauptsächlich Aufgaben mit Auswahl- (*multiple choice*) oder Kurzantworten eingesetzt. Obschon man auch mit diesem Prüfungsformat gründliches Verstehen und komplexes Problemlösen erfassen kann (Wolf, 1994), zielen die meisten Fragen in standardisierten Schulleistungstests auf isoliertes Faktenwissen und Routineverfahren (Resnick & Resnick, 1992). Sie beziehen sich damit auf zwei Komponenten der Fachleistungen, die sich leicht operationalisieren und zweifelsfrei beurteilen lassen (Shepard, 1992), unter lebensweltlichen Gesichtspunkten aber nur ein lückenhaftes Bild der Leistungsfähigkeit des Testnehmers abgeben. Diese Schwerpunktsetzung kann sich ungünstig auf die Unterrichtsgestaltung auswirken. Um das Ansehen der betreffenden Schule durch gute Testleistungen zu heben, lassen manche Lehrpersonen ihre Schülerinnen und Schüler über Gebühr Faktenwissen auswendig lernen und Routineverfahren üben. Dadurch verkürzen sie die zur Verfügung stehende Zeit zur Förderung anderer ebenso wichtiger Kompetenzen wie gründliches Verstehen, vernetztes Denken, kreatives Problemlösen, sachbezogenes Argumentieren und kompetente Selbststeuerung. Ausserdem schöpfen sie ihr didaktisches Repertoire nicht aus.

Wachgerüttelt durch die unliebsamen Wirkungen der staatlich verordneten, vom Behaviorismus geprägten Schulleistungstests (Gage & Berliner, 1992) auf die individuellen Kompetenzen der Testnehmer und die schulischen Lehr-Lern-Kulturen, fordern Kritiker in den USA seit einigen Jahren Messinstrumente, die dem aktuellen Verständnis von Lehren und Lernen, bzw. der generellen Entwicklung in der Lehr-

Lern-Forschung besser entsprechen: Seit der kognitiven Wende in der Psychologie (Miller, Galanter & Pribram, 1960/1973) wird Lernen primär als aktive, konstruktive, kumulative und zielgerichtete Tätigkeit aufgefasst. In den Achtzigerjahren haben Ergebnisse der Expertise-, der Kooperations- und der Metakognitionsforschung eindrücklich gezeigt, dass Lernen ebensosehr eine situierte, interaktive und selbstgesteuerte Tätigkeit ist. Auf dieser Grundlage ist ein erweiterter kognitionspsychologischer Lernbegriff entstanden (Stebler, Reusser & Pauli, 1994). Parallel dazu ist der Ruf nach *authentischen Lehr-Lern-Umgebungen* laut geworden. Man hat dabei didaktische Arrangements vor Augen, die bezüglich Sinnhaftigkeit, Interaktionskultur, Kompetenzspektrum, Wissensaufbau und Lernsteuerung grosse Ähnlichkeit mit den Arbeitsgemeinschaften der Erwachsenen haben. In authentischen Lehr-Lern-Umgebungen bearbeiten die Schülerinnen und Schüler alleine und/oder im Team fachlich relevante und anspruchsvolle Probleme, kommunizieren miteinander, helfen einander, nutzen ausserschulische Informationsquellen, übernehmen schrittweise Verantwortung für ihr Tun und planen, überwachen und beurteilen ihr Vorgehen zunehmend selbst (Collins, Brown & Newman, 1989). Um jene Kompetenzen verlässlich zu messen und zu beurteilen, die in ausserschulischen Lebens- und Arbeitsgemeinschaften verlangt und in sogenannt authentischen Lehr-Lern-Umgebungen gefördert werden, braucht es Testinstrumente, mit denen sich Fachleistungen und Lernkompetenzen kontextualisiert und umfassend messen lassen. Will man beispielsweise Aussagen über 'authentisches' Problemlösen machen, so sind bei der Leistungsmessung u.a. Attribute zu berücksichtigen, wie sie in Abbildung 1 dargestellt sind. Es gilt die Grundregel, dass das Testarrangement dem 'authentischen' Lern- und Anwendungskontext möglichst gut entsprechen sollte. Wenn dies der Fall ist, haben Verfahren und Ergebnisse der Leistungsmessung und -beurteilung eine gewisse Authentizität. Man spricht dann von *authentic assessment* (Wiggins, 1989a).

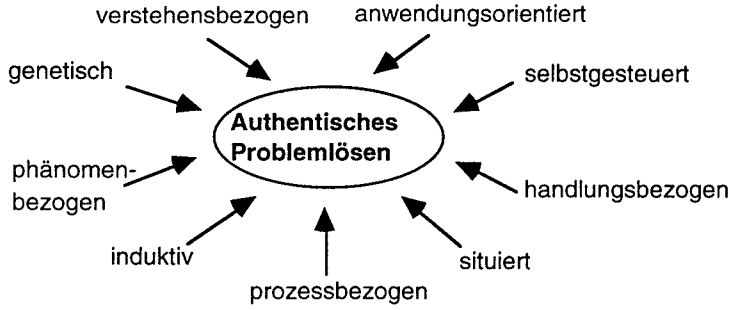

Abbildung 1: Einige Facetten des schillernden Begriffs *authentisches Problemlösen.*

Unter dem Leitmotiv "Creating tests worth taking" (Wiggins, 1992) fordern die Testkritiker in den USA höhere Authentizität beim Prüfen. Sie wollen damit zwei Ziele erreichen. Erstens sollen Fachleistungen und Lernkompetenzen in komplexen Anwendungssituationen gemessen und beurteilt werden. Solche Situationen haben vier Kennzeichen (Wiggins, 1989b):

(1) Sie repräsentieren relevante Bildungsinhalte.
(2) Die Kriterien der Leistungsbeurteilung sind transparent.
(3) Die Selbstbeurteilung des Testnehmers wird mitberücksichtigt.
(4) Der Testnehmer macht seine Lernergebnisse und Lernwege publik.
Zweitens soll über den Mechanismus "what you test is what you get" (Silver, 1992) versucht werden, die Unterrichtsgestaltung und die Lern- und Denkprozesse der Schülerinnen und Schüler wieder stärker in Richtung Verstehen, Problemlösen und Selbstreflexion zu lenken (Shepard, 1992).

2 Experimentieraufgabe und Testportfolio - zwei 'authentische' Anwendungssituationen

Experimentieraufgaben und Testportfolios sind Methoden der Leistungsmessung und -beurteilung, denen in erziehungswissenschaftlichen Publikationen das Zertifikat 'authentisch' verliehen wird (Linn, Baker & Dunbar, 1991). Mit beiden Arrangements können sowohl Fachleistungen als auch Lernkompetenzen erfasst werden. Die Settings unterscheiden sich aber in Bezug auf die Art und die Dauer der lernerbezogenen Datenerhebung. Mit Experimentieraufgaben werden in separaten Sitzungen punktuelle Messungen vorgenommen. Testportfolios sind Sammlungen von Lernspuren, die im regulären Unterricht über Wochen oder Monate hinweg entstanden sind.

2.1 Experimentieraufgaben - fachlich und strategisch anspruchsvolle Handlungsprobleme

Experimentieraufgaben (Hands-on Performance) sind komplexe, fachlich verankerte *Handlungsprobleme*, die relevante Bildungsinhalte verkörpern, verschiedene Lösungswege zulassen und strategisches Verhalten auslösen. Je nach Aufgabe müssen die Lernenden ein greifbares Produkt herstellen oder Objekte bezüglich bestimmter Merkmale vergleichen, die Bestandteile eines Ganzen identifizieren, Objekte klassifizieren, einen Vorgang systematisch beobachten und protokollieren sowie durch logisches Denken und vernünftiges Urteilen begründete Schlussfolgerungen ziehen (Shavelson, Solano-Flores & Ruiz-Primo, 1997). Es wird erwartet, dass sie sich wie Forschende verhalten, die Hypothesen generieren, planen, prüfen, überwachen, Daten protokollieren, beurteilen und die Ergebnisse anhand ihres Fachwissens erklären.

Im Rahmen eines weltumspannenden Schulleistungstests, des TIMSS[1] Performance Assessment, wurden auch in 44 Deutschschweizer Klassen der siebten Jahrgangsstufe (N=396) mathematische und naturwissenschaftliche Handlungsprobleme zur Leistungsmessung eingesetzt. Die Erhebung wurde in Form eines Postenlaufs bzw. Circuits durchgeführt. Die Jugendlichen mussten Experimente planen und durchführen, Ergebnisse protokollieren, Trends aus den Daten herauslesen, Schlussfolgerungen ziehen und diese mit Hilfe ihres mathematischen und naturwissenschaftlichen Fachwissens erklären. In einzelnen Fällen war auch eine abschliessende Beurteilung des Experimentierverhaltens verlangt. Die Deutschschweizer Siebtklässler haben trotz der Tatsache, dass sie zum Testzeitpunkt aufgrund des späten Einschulungstermines ein Schuljahr weniger absolviert hatten als die Probanden in den meisten anderen Ländern, im internationalen Vergleich ausgezeichnet abgeschnitten (Harmon et al., 1997; Stebler, Reusser & Ramseier, 1997, 1998). Deutlich höhere Testwerte erzielten nur die Achtklässler aus Singapur. Beim Problemlösen lagen die Testwerte der Deutschschweizer Jugendlichen näher bei jenen von Singapur und deutlicher über dem internationalen Mittelwert als beim Anwenden von Routineverfahren. Die Leistungen der Knaben und der Mädchen unterschieden sich nicht.

[1] TIMSS ist das Kürzel für Third International Mathematics and Science Study.

Die Experimentieraufgabe *Puls* ist ein komplexes Handlungsproblem, das die Schülerinnen und Schüler beim TIMSS Performance Asessment lösen mussten. Es wurde den Testpersonen als halbstrukturiertes Schülerexperiment mit drei Teilaufgaben vorgelegt (Abb. 2). Teilaufgabe 1 verlangt einen Versuchsplan (planen), mehrere Messungen (durchführen) und ein Datenprotokoll (protokollieren). Beim Planen müssen Entscheidungen über die Anzahl, die Abstände und die Dauer der Messungen gefällt werden. Ein aussagekräftiges Experiment erfordert eine Messung im Ruhezustand und mindestens vier Messungen nach körperlicher Aktivität. Beim Durchführen des Experimentes wechseln Phasen der Gymnastik mit Phasen des Pulsmessens und Protokollierens ab. Was sich so einfach liest, ist in Wirklichkeit ein anspruchsvolles strategisches Verhalten, bei dem der Puls ertastet und gezählt, die Uhrzeit abgelesen und eine fehleranfällige Multiplikation (Pulsschläge mal Minutenbruchteil) ausgeführt werden muss. Beim Protokollieren der Messergebnisse müssen die Spalten der Tabelle mit Titeln versehen, die Messwerte (Pulsfrequenz) den Messzeitpunkten zugeordnet und die Dimensionen (Min., Sek.) richtig eingetragen werden. Die so entstandene Tabelle ist Voraussetzung für Teilaufgabe 2, bei der die Testperson einen Trend aus den Daten herauslesen und beschreiben muss. Teilaufgabe 3 verlangt eine Interpretation der Untersuchungsergebnisse. Erwartet wird die Erklärung, dass der Puls bei körperlicher Anstrengung deshalb steigt, weil die Muskeln mehr Sauerstoff brauchen und diese Zufuhr durch die Erhöhung des Herzminutenvolumens erreicht wird.

Bei TIMSS wurde der Test als *Circuit* durchgeführt. Die Probanden besuchten nach einem Rotationsplan mehrere Posten. An jedem Posten fanden sie das Material und eine schriftliche Experimentieranleitung, die gleichzeitig als Protokollblatt diente. Je nach Komplexität der Experimentieraufgabe hatten sie ein Zeitbudget von 15 oder 30 Minuten. Für das Handlungsproblem 'Puls' waren 15 Minuten vorgesehen.

Die *Beurteilung* der Schülerantworten erfolgte anhand eines inhaltsanalytischen Kategoriensystems. Für jede Teilaufgabe wurden Kriterien und Kategorien formuliert. Es wurde ein zweistelliger Code verwendet. Durch die erste Ziffer wurde die Qualität (z.B. ansatzweise, teilweise oder ganz richtig), durch die zweite die Spezifität (z.B. Lösung mit Fachbegriff, Lösung mit Alltagswissen) der Antwort erfasst (Tabelle 1).

Wenn man die Punktwerte für die Lösungsqualität der Teilaufgaben pro Handlungsproblem addiert, gelangt man zu Aussagen über die Fachleistungen. Zählt man über verschiedene Experimentieraufgaben hinweg beispielsweise die Punktwerte jener Teilaufgaben zusammen, bei denen der Testnehmer planen musste, erhält man Informationen über einen Aspekt seiner Lern- oder Problemlösekompetenzen. Die Codes für die Spezifität der Schülerantworten geben Aufschluss über das Spektrum der Lösungsansätze und die Streuung der Fehlertypen in der Teststichprobe.

PULS

An dieser Station brauchst Du:
eine Uhr mit einem Sekundenzeiger
eine Stufe am Boden, um hinaufzusteigen

Lies ALLE Anweisungen genau durch!

Deine Aufgabe:
Finde heraus, wie sich Dein Puls verändert, wenn Du 5 Minuten lang die Stufe auf- und absteigst.

Folgendes sollst Du tun:

- Fühle Deinen Puls und vergewissere Dich, dass Du weisst, wie man ihn zählt. WENN DU DEINEN PULS NICHT FINDEN KANNST, BITTE EINEN TEST-LEITER UM HILFE.
- Entscheide, wie oft Du Messungen durchführen möchtest. Zähle Deinen Puls zum ersten Mal, bevor Du mit dem Auf- und Absteigen beginnst.
- Steige ungefähr 5 Minuten lang die Stufe auf und ab, und miss Deinen Puls in regelmässigen Abständen.

1. Erstelle eine Tabelle, in der Du die Zeiten aufschreibst, zu denen Du Deinen Puls gemessen hast sowie die Messwerte, die Du erhalten hast.

Start		**-->**	**10"**	**----------**	**17**
nach	**1 min**	**-->**	**10"**	**----------**	**21**
	2 min	**-->**	**10"**	**----------**	**24**
	3 min	**-->**	**10"**	**----------**	**26**
	4 min	**-->**	**10"**	**----------**	**30**
	5 min	**-->**	**10"**	**----------**	**30**

2. Wie veränderte sich Dein Puls während dieser Turnübung?

Er hat kontinuierlich zugenommen. Nach 5 min war der Puls fast doppelt so schnell, wie zu Beginn.

3. Warum hat sich Dein Puls auf diese Weise verändert?

Damit meine Muskeln arbeiten konnten, brauchten sie Sauerstoff. Um diesen herbeizuführen, musste mein Herz mehr mit Sauerstoff versorgtes Blut in den Umlauf bringen. Dafür musste es schneller arbeiten und damit schlagen.

Räume Deinen Arbeitsplatz so auf, wie Du ihn vorgefunden hast, damit diese Station wieder benutzt werden kann!

Abbildung 2: Testheft der Experimentieraufgabe *Puls* mit Schülerlösung. Das Originaltestheft bestand aus zwei A4 Seiten.

Tabelle 1: Kriterien und Kategorien zur Beurteilung des Aspektes *Qualität der Daten* bei der ersten Teilaufgabe des Handlungsproblems 'Puls'.

CODE	ANTWORT
Vollständige Antwort	
30	Die Pulsfrequenz liegt innerhalb der Bandbreite und steigt vom Anfang bis zum Ende der Tabelle an.
31	Die Pulsfrequenz liegt innerhalb der Bandbreite, nimmt kontinuierlich zu, pendelt sich auf hohem Niveau ein oder sinkt leicht.
39	Andere vollständige Antworten.
Teilweise richtige Antwort	
20	Weniger als 5 Messungen: die übrigen Kriterien sind erfüllt.
29	Andere teilweise richtige Antworten, die zwei Kriterien erfüllen.
Minimal richtige Antwort	
10	Vollständige Eintragungen mit kleinen Fehlern: z.B. eine od. zwei Messungen sind mit den restlichen nicht vereinbar, aber es sind genug Daten vorhanden, um den allgemeinen Trend zu erkennen.
11	Nur die Messung zu Beginn und am Schluss wird festgehalten.
12	Die Beschreibung der Pulsfrequenz erfolgt qualitativ anstatt quantitativ. Ein allgemeiner Trend ist erkennbar. Beispiel: langsam, mittel, schnell oder auf/ab.
19	Andere minimal richtige Antworten.
Falsche Antwort	
70	Keine Zeitmessungen eingetragen.
71	Die Pulsfrequenz ist nicht reell, weil entweder die Zeit oder die Pulsfrequenz oder beides unglaubwürdig ist. Beispiel: Der Schüler beschreibt den Puls als ganze Zahlen pro Sekunde.
79	Andere falsche Antworten.
Keine Antwort	
90	Durchstrichen/ausradiert, nicht lesbar, nicht interpretierbar.
99	Leer.

Kriterien für eine vollständige Antwort:

I) Mindestens 5 Messungen sind in einer Tabelle eingetragen; und zwar eine für den Ruhezustand und mindestens 4 während der Übung.
II) Die Pulsfrequenz liegt innerhalb einer Spanne von 7 bis 30 Schlägen pro 10 Sekunden (d.h. 40 bis 210 Pulsschläge pro Minute).
III) Die Pulsfrequenz steigt während der Übung; sie kann gegen das Ende hin aber auch auf gleichem Niveau verharren oder leicht sinken.

2.2 *Testportfolios - lernzielbezogene und kommentierte Sammlungen von Lernspuren*

Wenn sich junge Künstlerinnen oder Künstler an einer Akademie bewerben, legen sie seit jeher ein Portfolio vor. Diese Mappe enthält ihre besten Arbeiten und zeigt die Bandbreite ihres Könnens. Die Fachjury beurteilt die Werke nach bestimmten Kriterien und entscheidet anhand des Ergebnisses sowie einer Aufnahmeprüfung und allfälliger Empfehlungsschreiben, ob ein Anwärter den begehrten Studienplatz erhält oder abgewiesen wird. Was bei Künstlern gang und gäbe ist, könnte sich auch in der

Schule bewähren. Im angelsächsischen Raum werden Portfolios schon seit einiger Zeit zur formativen und summativen Beurteilung von Schulleistungen eingesetzt. Die Lehrpersonen werden via praktische Anleitungen mit dieser Testmethode vertraut gemacht (Paulson, Paulson & Meyer, 1991; Stenmark, 1991; Vavrus, 1990). Hierzulande gibt es erste Erfahrungen mit Testportfolios in der Grund- und Weiterbildung von Lehrkräften an Berufsschulen (Behrens, 1997). Ansonsten wird diese Form der Leistungsbeurteilung an Schweizer Schulen noch kaum systematisch praktiziert.Wir halten uns in der Fortsetzung daher an die Anleitung zum Einsatz von Testportfolios in Mathematik, die Stenmark (1991) für Lehrpersonen in den USA verfasst hat.

In didaktischen Wegleitungen wird zwischen Arbeits- und Testportfolios unterschieden (Stenmark, 1991). Arbeitsportfolios sind im weitesten Sinne Ordner oder Mappen, in denen die Schülerinnen und Schüler fertige und laufende Arbeiten systematisch ablegen. Sie bilden die Materialsammlung, aus der periodisch Testportfolios zusammengestellt werden. Ein *Testportfolio* soll Auskunft geben über die Leistungsentwicklung der Inhaberin oder des Inhabers in einem bestimmten Zeitraum. Es enthält Informationen über Fachleistungen und Lernkompetenzen. Bezüglich welcher Aspekte der Fachleistungen und Lernkompetenzen diese Entwicklung dokumentiert und beurteilt wird, hängt von den Zielen des Unterrichts in der Testperiode ab. Diese Ziele wiederum richten sich nach den Vorgaben des Lehrplanes. Abbildung 3, die wir in Anlehnung an Stenmark (1991) erstellt haben, gibt einen Eindruck der Schülerdokumente, die ein Testportfolio zum Thema 'Kreisberechnungen' enthalten könnte. Es soll Auskunft geben über die Fachleistungen in Bezug auf die Lernziele 'Kenntnisse und Fertigkeiten' und 'Problemlösen und Verstehen' sowie über die Lernkompetenzen in den Bereichen 'Lernstrategien' und 'Selbstbeurteilung'. Wenn aufeinanderfolgende Testportfolios zu verschiedenen Themen auf ähnliche Lernziele bezogen sind, kann die Leistungsentwicklung nicht nur innerhalb einer Testperiode, sondern auch über längere Zeiträume hinweg systematisch dokumentiert werden.

Es ist davon auszugehen, dass im traditionellen Unterricht zahlreiche Dokumente zu Fachleistungen, aber nur wenige zu Lernkompetenzen anfallen, sofern die Lehrperson keine entsprechende Reflexionskultur aufbaut. Wir wollen daher anhand eines Beispiels andeuten, welche Informationen über metakognitive Aktivitäten bzw. verschiedene Aspekte der Lernsteuerung sich durch die Analyse von Lern- und Reisetagebüchern (Ruf & Gallin, 1999), Arbeitsblättern mit Reflexionsspalten (Beck, Guldimann & Zutavern, 1991), schriftlichen Arbeitsrückblicken oder Kurzberichten über Prüfungsvorbereitungen gewinnen lassen. Die folgenden, in der Originalschreibweise abgedruckten Texte, haben wir in einer Studie mit Schülerinnen und Schülern der siebten Jahrgangsstufe eingeholt .

- "Für Mathe lerne ich nicht viel, weil ich es im grossen und ganzen schnalle. Wenn nicht dann frage ich meinen Vater, auch wenn er nicht weiss um was es geht. Dabei finde ich ganz plötzlich von alleine die lösung. Das passiert mir oft in Geometri. Aber eigendlich lerne ich nicht in Mathe weil ich kann ja nicht die Rechnungen auswendig lernen, und die Regeln und Theorie kann ich mir immer schon in der Stunde merken.
- Ich fange erst etwa 3-4 Tage vor der Prüfung an zu üben. Dann schaue ich im Heft nach, wo ich viele Fehler hatte und gehe diese Rechnungen im Buch noch mal durch. Wenn ich dann immer noch nicht drauss komme, frage ich den Lehrer. Im Heft schreiben wir oft Theorien ein. Diese übe ich meistens so etwa 2 Tage vor der Prüfung. All das tue ich nur um eine möglichst gute Note zu bekommen.

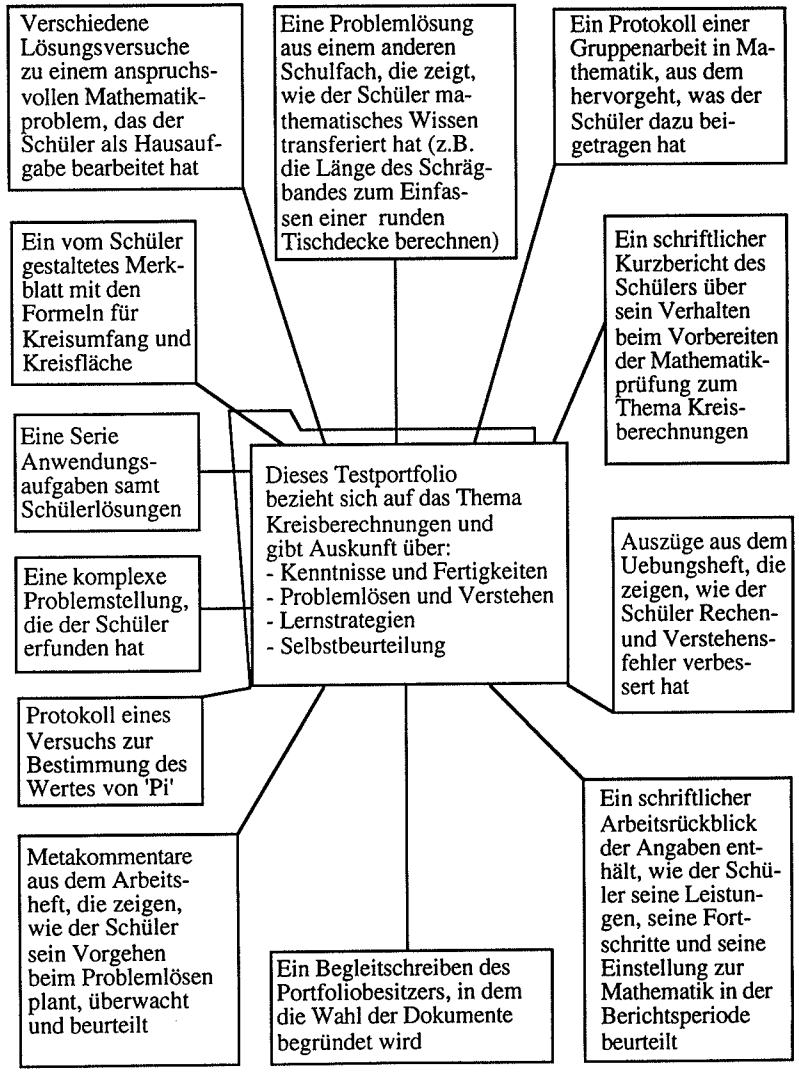

Abbildung 3: Dokumente, die ein Testportfolio in Mathematik enthalten könnte (in Anlehnung an Stenmark, 1991).

• Mein Vorbereiten beginnt schon viel früher. Jedesmal, wenn ich eine Aufgabe, Erklärung usw. nicht vestehe, drehe ich das heft um und schreibe dort die Aufgabe und Lösung auf. Wenn ich dann zu arbeiten beginne, weiss ich genau was ich schon kann, was nicht. Das lerne ich dann auswendig und versuche es zu verstehen. Das andere schaue ich nur kurz an.
• Im Franzosisch habe ich grosse mühe. Wenn ich für eine Prüfung lerne, repetiere ich den ganzen Stoff, der in der Prüfung gefragt wird. Aber meistens lerne ich zuwenig, oder das falsche. Ich glaube, dass ich mich zu schnell ablenken lasse. Meistens verschiebe ich das Lernen bis zum letzten Tag und dann stehe ich unter Zeitdruck.
• Ich schau meistens im Buch nach oder im Ecrit und versuche Beispiele zu diesem Thema zu machen. Aber da ich meistens die hälfte nicht richtig gutt behersche gehe ich zu meinem Grossvater. Und er erklärt mir dann alles noch einmal. Wenn ich Wörter lernen muss schreibe ich sie zuerst einmal ab, dann decke ich erst das Deutsche ab. Wenn ich das einigermassen kann decke ich das Französische ab. Danach lasse ich mir von meiner Mutter die Wörter diktieren. Das heisst sie sagt es auf Deutsch und ich sage und schreibe es auf Französisch."

Solche und ähnliche Texte können Aufschluss darüber geben, wie gründlich Lernende über ihr Denk-Handeln nachdenken, wie sie ihr Verhalten planen, überwachen und beurteilen, welche Strategien sie dabei einsetzen, welche Schwierigkeiten sie erleben, welche Motive ihr Verhalten leiten, wie differenziert ihre Begriffe für den Bereich Denken und Lernen sind und wie präzis sie über kognitive Aktivitäten und Befindlichkeiten berichten. Entsprechende Informationen sind nicht nur mit Bezug auf eine umfassende Leistungsbeurteilung, sondern auch im Hinblick auf eine gute Lernberatung unerlässlich.

Eine Lehrperson, die von ihren Schülerinnen und Schülern ein Testportfolio verlangt, muss in Abhängigkeit vom Ziel der Leistungsbeurteilung vorab vier Fragen klären:

(1) Soll das Testportfolio nur die besten Arbeiten der Schülerin oder des Schülers enthalten oder die Leistungsvarianz dokumentieren?
(2) Sollen nur fertige Arbeiten oder auch Entwürfe einbezogen werden?
(3) Zu welchen fachlichen und überfachlichen Lernzielen sollen Informationen gesammelt werden?
(4) Nach welchen Kriterien werden die Dokumente beurteilt?

Um einen Eindruck von lernzielbezogenen Kriterien zur Beurteilung von Testportfolios in Mathematik zu vermitteln, präsentieren wir eine aus dem Englischen übersetzte und leicht modifizierte Kriterienliste von Stenmark (1991).

- Versteht die Aufgaben und Problemstellungen.
- Wendet verschiedene Lösungsansätze und Strategien an.
- Sucht und findet Lösungen für neuartige und anspruchsvolle Probleme.
- Setzt geeignete Hilfsmittel ein.
- Baut mathematisches Wissen auf, indem er/sie Sachverhalte erforscht, Beziehungen herstellt, Verfahren erfindet sowie mathematische Aussagen und Vorgehensweisen hinterfragt.
- Verwendet die mathematische Fach- und Symbolsprache richtig.
- Verifiziert die Lösungen und interpretiert die Ergebnisse.
- Bringt sich beim Lernen und Problemlösen in Paaren und Kleingruppen ein.
- Setzt Mathematik in Beziehung zu anderen Fächern und zur Alltagserfahrung.
- Hat eine positive Einstellung zur Mathematik (Vertrauen in die eigenen Fähigkeiten, Beweglichkeit, Ausdauer, Wertschätzung der Mathematik).
- Beurteilt eigenes Vorgehen und eigene Lösungen und nimmt gegebenenfalls Korrekturen vor.

Der Autor weist einleitend darauf hin, dass seine Zusammenstellung als Handreichung für Lehrpersonen gedacht ist und keinen Anspruch auf Vollständigkeit erhebt. Er erachtet es als selbstverständlich, dass jede Lehrperson anhand des Curriculums und der Unterrichtsziele eine eigene Kriterienliste zusammenstellt, aus der sie bei der Beurteilung eines Testportfolios, das einen Zeitraum von zwei bis drei Wochen abdeckt, etwa drei Kriterien auswählt. Diese Kriterien werden den Schülerinnen und Schülern an konkreten Beispielen erläutert und ausgehändigt. Die Schüler lassen sich bei der Auswahl der Dokumente von den *Beurteilungskriterien* leiten.

Ein Testportfolio kann sukzessive über Tage und Wochen entstehen oder im Hinblick auf eine bevorstehende Leistungsbeurteilung in einem Zuge zusammengestellt werden. In beiden Fällen wählt die Schülerin die Dokumente, die sie nach einem bestimmten System im Testportfolio ablegen will. Da im Testportfolio Entwicklungsverläufe mit Bezug auf Fachleistungen und Lernkompetenzen dokumentiert werden sollen, ist es entscheidend, dass alle Dokumente datiert sind. Bei Bedarf ergänzt die Lehrperson das Portfolio. Die Schülerin erstellt ein Inhaltsverzeichnis, in dem sie die Wahl der Dokumente begründet. Anstelle eines kommentierten Inhaltsverzeichnisses kann auch eine Lesehilfe in Form einer Zusammenfassung verlangt werden, in der die Schülerin neben den Begründungen für die Dokumentenwahl ihre Erwartungen in Bezug auf die bevorstehende Leistungsbeurteilung zum Ausdruck bringt.

Beim Beurteilen der Testportfolios stellt die Lehrperson zum einen fest, ob und inwiefern die formulierten Kriterien erreicht sind. Zum anderen versucht sie durch den Vergleich von Dokumenten, die früher oder später in der Testperiode entstanden sind, Informationen über den Entwicklungsverlauf hinsichtlich der gewählten Kriterien zu gewinnen. Ihre Eindrücke hält die Lehrperson in Form von Notizen oder mit Hilfe von Likert-Skalen (z.B. nicht erfüllt, ansatzweise erfüllt, mehrheitlich erfüllt, vollumfänglich erfüllt) fest. Dieses Datenmaterial fasst sie anschliessend in einem schriftlichen Kurzbericht zusammen, dessen Ausrichtung vom Ziel der Leistungsmessung (formativ vs. summativ) abhängt. Bei Bedarf kann sie ergänzend auch eine Note setzen. Diese Beurteilung wird ins betreffende Testportfolio gelegt und mit dem Inhaber oder der Inhaberin besprochen.

Das bislang skizzierte Vorgehen beim Gestalten und Beurteilen von Testportfolios kann je nach den Zielen der Leistungsmessung, den Adressaten (Lehrpersonen, Eltern, Behörden), den Voraussetzungen der Klasse, den Vorlieben der Lehrperson oder den einbezogenen Unterrichtsfächern variiert werden.

3 Authentizität beim Prüfen heisst verschiedene Testmethoden kombinieren

Im deutschen Sprachraum stehen wir weder vor dem Problem, valide Instrumente für schulübergreifende Leistungsvergleiche zu entwickeln, noch Mittel und Wege zu finden, um die unerwünschten Spätfolgen staatlich verordneter Schulleistungstests zu kurieren. Trotzdem lohnt sich das Nachdenken über Vorzüge und Fallstricke der Lei-

stungsmessung und -beurteilung in komplexen Anwendungssituationen im Hinblick auf die Frage, inwiefern entsprechende Testarrangements auch unsere Prüfungskultur, in der die Leistungsmessung primär im Dienste der Förder- und Selektionsdiagnostik steht, bereichern könnten. Ein Ausgangspunkt ist die Analyse der vorgestellten Prüfungsverfahren, Experimentieraufgaben und Testportfolio, anhand der vier einleitend genannten Kriterien:

Relevante Bildungsinhalte. Testverfahren, die das Zertifikat 'authentisch' verdienen, messen und beurteilen Fachleistungen und Lernkompetenzen in komplexen Anwendungssituationen. Damit berücksichtigen sie die lernpsychologische Erkenntnis, dass Lernen und Denken situierte Tätigkeiten sind. Sie beziehen sich auf relevante Bildungsinhalte und prüfen Kenntnisse und Fertigkeiten, die in qualitativ hochstehendem, verstehensorientiertem Unterricht aufgebaut und geübt werden. Die Anwendungssituationen sind für die Probanden interessant, bedeutsam und lehrreich. Diese Forderungen sind sowohl bei der Pulsaufgabe als auch beim Testportfolio erfüllt. Die Experimentieraufgabe 'Puls' repräsentiert ein Kernthema des Biologieunterrichts, setzt Einsicht in den Sachzusammenhang 'Sauerstofftransport des Blutes' voraus, verlangt Kenntnisse über die Struktur eines naturwissenschaftlichen Experimentes, prüft Denkprozesse im symbolischen, ikonischen und enaktiven Medium und erfordert ein hohes Mass an Selbststeuerung. Das Testportfolio bezieht sich auf ausgewählte Inhalte des Mathematiklehrplanes, der in den Volksschulen der Deutschschweiz auf die Entwicklung selbständigen Denkens und der Ausdrucksfähigkeit, die Förderung von Kreativität sowie auf die Vernetzung der Mathematik mit anderen Fächern ausgerichtet ist. Er setzt die Leitplanken für einen Unterricht, in dem das Vorstellungsvermögen erweitert und vertieft, Kenntnisse und Fertigkeiten erworben, die Mathematisierfähigkeit und das Problemlöseverhalten gefördert werden sollen. Das Zusammenstellen und Kommentieren des Testportfolios ist für die Schülerin oder den Schüler eine anspruchsvolle Problemlöseaufgabe, die ein ausgeklügeltes Zusammenspiel von kognitiven und metakognitiven Kompetenzen erfordert.

Die Kriterien der Leistungsbeurteilung sind transparent. Die Leistungsbeurteilung erfolgt nach Kriterien, welche dem Testnehmer mitgeteilt werden. Durch die Transparenz der Beurteilungskriterien kann das Lernen eine stärkere Zielorientierung erhalten. Diese Forderung ist bei der Pulsaufgabe nur teilweise erfüllt. Hier werden für jede Teilaufgabe inhaltliche Kriterien formuliert und darauf bezogene Kategorien entwickelt. Diese werden jedoch erst nach der Beurteilung offengelegt. Beim Testportfolio hingegen werden dem Probanden die Kriterien zusammen mit dem Prüfungsauftrag ausgehändigt. Er kann sich beim Zusammenstellen und Kommentieren des Portfolios davon leiten lassen und die Rückmeldungen der Lehrperson damit vergleichen.

Die Selbstbeurteilung des Testnehmers wird mitberücksichtigt. Die Selbstbeurteilung wird stärker thematisiert als bei herkömmlichen Testverfahren. Bei der Pulsaufgabe wird diese Forderung nur implizit realisiert, nämlich dadurch, dass der Testnehmer bei seinem strategischen Verhalten sein Vorgehen und die Ergebnisse laufend mit seinen Zielen und Gütestandards vergleicht und die resultierenden Informationen zur Steuerung seines Verhaltens nutzt. Es gibt jedoch Experimentieraufgaben, bei denen explizit eine abschliessende Beurteilung und allfällige Änderungsvorschläge verlangt werden. Beim Zusammenstellen von Testportfolios wird durch die Auflage,

die Dokumentenwahl zu begründen, ergänzend zur impliziten auch eine explizite Selbstbeurteilung verlangt. Dies stellt hohe Anforderungen an die Schülerinnen und Schüler, denen die meisten erst nach längeren Aufbauphasen gewachsen sind. Im Hinblick auf verlässliche Selbstbeurteilungen müssen die Schülerinnen und Schüler zum einen wissen, was beurteilt wird, bzw. was sie beurteilen müssen. Zum anderen brauchen sie präzise Gütestandards. Ohne eine hinreichend genaue Vorstellung davon, was beipielsweise eine originelle Lösung, eine elegante Formulierung, eine saubere Darstellung oder ein angemessener Aufwand ist, ist keine hinreichende Übereinstimmung zwischen Fremd- und Selbstbeurteilung zu erwarten. Es ist daher in jedem Fall entscheidend, dass die Kriterien der Leistungsbeurteilung mit den Schülerinnen und Schülern gründlich und an Beispielen erörtert werden. Zu bedenken ist ferner, dass man mit der Aufforderung zur Selbstbeurteilung in die Privatsphäre der Testnehmerin oder des Testnehmers eindringt. Dies kann bei ungenügender Vertrauensbasis vor allem für unsichere Schülerinnen und Schüler sehr belastend sein.

Der Testnehmer macht seine Lernergebnisse und Lernwege publik. Die Testnehmer machen ihr Denken und Handeln anderen Personen zugänglich, indem sie ausführlich darüber sprechen oder schreiben. Auf diese Weise werden intrapersonale Prozesse auf die soziale Ebene verlegt. Sowohl bei der Pulsaufgabe als auch beim Testportfolio geschieht dieses Externalisieren in schriftlicher Form. Es ist aber durchaus denkbar, dass Arrangements geschaffen werden, in denen die Probanden ihre Untersuchungsergebnisse oder Testportfolios mündlich vorstellen und anschliessend mit dem Testleiter oder der Klasse diskutieren.

Wie diese Analyseergebnisse zeigen, tragen Testarrangements wie Experimentieraufgaben und Portfolios insbesondere der situierten, zielbezogenen, selbstgesteuerten und interaktiven Natur des Lernens und Problemlösens besser Rechnung als traditionelle Schulleistungstests im Mehrfach- oder Kurzantwortformat. Sie ermöglichen dadurch eine umfassendere Beurteilung von Fachleistungen und Lernkompetenzen und unter förderdiagnostischen Gesichtspunkten eine differenziertere Lernberatung. Neben diesen Vorzügen bergen sie aber auch messtechnische Fallstricke, insbesondere hinsichtlich der *Validität* (Linn, Baker und Dunbar, 1991). Wer stark kontextualisierte Messverfahren zur Selektion oder bei klassenübergreifenden Leistungsvergleichen einsetzen will, sollte auch folgende Punkte beachten:

Die Generalisierbarkeit der Testergebnisse ist limitiert. Die Leistungen in komplexen Anwendungssituationen hängen stark von der je spezifischen Problemstellung und dem Testzeitpunkt ab. Es lassen sich daraus nur bedingt Schlüsse über Leistungen in anderen Fachbereichen, bei anderen Themen und Aufgabentypen ableiten. An dieser Tatsache lässt sich auch durch die Kombination verschiedener Anwendungsaufgaben kaum rütteln, denn aufgrund der vergleichsweise langen Bearbeitungszeit können den Probanden pro Testsitzung nur wenige Aufgaben vorgelegt werden.

Das Spektrum der Lernkompetenzen muss abgedeckt sein. Im Gegensatz zu herkömmlichen Schulleistungstests, wo es vor allem darauf ankommt, dass eine relevante Auswahl zentraler Fachkonzepte abgedeckt ist, haben Leistungsmessungen in komplexen Anwendungssituationen zum Ziel, auch Lernkompetenzen zu erfassen. Im Hinblick auf aussagekräftige Ergebnisse ist darauf zu achten, dass die für den ent-

sprechenden Sachbereich relevanten Lern- und Problemlösestrategien in der Testsituation auch tatsächlich verlangt sind und in geeigneter Form erfasst werden.

Die Anwendungssituationen müssen für alle Testnehmer fair sein. Wenn Leistungen in komplexen Anwendungssituationen gemessen werden, kann die situative Einbettung des erforderlichen Sach- und Problemlösewissens einen starken Einfluss auf die Testergebnisse haben. Es konnte wiederholt gezeigt werden, dass Testpersonen dasselbe Lösungsverfahren in einem Kontext richtig anwenden, während sie sich in einem anderen Kontext nicht einmal daran erinnern. Angesichts dieser Tatsache ist eine angemessene thematische Vielfalt der Testaufgaben unabdingbar, um die Voraussetzungen der Knaben und der Mädchen sowie der Testnehmer verschiedener Ethnien ausgewogen zu berücksichtigen.

Abklären, ob Aufwand und Ertrag in einem vertretbaren Verhältnis stehen. Leistungsmessungen in komplexen Anwendungssituationen sind mit Bezug auf die Instrumentenentwicklung, die Testdurchführung, die Auswertung und die Berichterstattung sehr personal- und zeitaufwendig im Vergleich zu standardisierten Schulleistungstests. Dem Entscheid für solche Testarrangements muss daher die Frage vorausgehen, ob man dieselben Fachleistungen und Lernkompetenzen mit ökonomischeren Methoden nicht ebenso verlässlich messen und beurteilen könnte.

Die vorliegende Erörterung der Vorzüge und Fallstricke der Leistungsmessung und -beurteilung in komplexen Anwendungssituationen legt nahe, dass es sinnvoll ist, beim Beurteilen von Schulleistungen mehr als eine Methode einzusetzen und das Prüfarrangement gut auf die Lernziele abzustimmen. Dies ist schon deshalb wichtig, weil auch in verstehensorientiertem Unterricht das Fachwissen und die Routinen nicht vernachlässigt werden dürfen. Expertise beruht auf einem reichhaltigen, gut vernetzten Sachwissen und eingeschliffenen Automatismen, die dem Denken die Datenbasis und den Freiraum für komplexe Verstehens- und Problemlösetätigkeiten verschaffen. Faktenwissen und Routinen können auch mit herkömmlichen Leistungstests verlässlich gemessen werden. Wenn es hingegen darum geht, die Einsicht in Sachzusammenhänge, den Aufbau folgerichtiger Argumentationsgänge oder die Lernsteuerung zu beurteilen, sind komplexe Anwendungsaufgaben oft besser geeignet. Durch die *Kombination verschiedener Testverfahren* lassen sich Fachleistungen und Lernkompetenzen umfassend beurteilen und Ergebnisse erzielen, die unter formativen und summativen Gesichtspunkten Authentizität beanspruchen können.

Abschliessend möchten wir darauf hinweisen, dass wir dem Wort *Authentizität* in Lern- und Prüfungskontexten mit einer gewissen Skepsis begegnen. Welche Lebens- und Lernsituationen man als authentisch, glaubwürdig oder echt bezeichnet, ist eine Definitionsfrage, die sehr unterschiedlich beantwortet werden kann. Da eine verbindliche, nicht zirkuläre Definition fehlt, sagt das Attribut 'authentisch' im Prinzip einzig aus, dass zwischen Test- und Lernsituation eine starke Beziehung besteht. In diesem Sinne kann auch ein reiner Wissenstest authentisch sein für eine reproduktionslastige Lehr-Lern-Kultur. Trotz dieser Kritik unterstützen wir das Anliegen, das mit dem Schlagwort 'authentic assessment' zum Ausdruck gebracht wird, nämlich die Forderung und das Bemühen, Fachleistungen und Lernkompetenzen vermehrt an komplexen Anwendungssituationen zu erwerben und zu beurteilen.

Literatur

Beck, E., Guldimann, T. & Zutavern, M. (1991). Eigenständig lernende Schülerinnen und Schüler. *Zeitschrift für Pädagogik, 37*, 735-768.

Behrens, M. (1997). Das Portfolio zwischen formativer und summativer Bewertung. *Beiträge zur Lehrerbildung, 15*(2), 176-184.

Collins, A., Brown, J. S. & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning and instruction. Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gage, N. L. & Berliner, D. C. (1992). *Educational Psychology* (5 ed.). Boston: Houghton Mifflin.

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., Gonzales, E. J. & Orpwood, G. (1997). *Performance Assessment in IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Linn, R. L., Baker, E. L. & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Miller, N. E., Galanter, E., & Pribram, K. H. (1973). *Strategien des Handelns. Pläne und Strukturen des Verhalten*. Stuttgart: Klett (Original 1960).

Paulson, F. L., Paulson, P. R. & Meyer, C. A. (1991). What makes a portfolio a portfolio? *Educational Leadership, 48*(5), 60-63.

Resnick, L. B. & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments. Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer.

Ruf, U. & Gallin, P. (1999). *Dialogisches Lernen in Sprache und Mathematik*. Bd. 2. Spuren legen - Spuren lesen. Unterricht mit Kernideen und Reisetagebüchern. Seelze-Velber: Kallmeyer.

Ruiz-Primo, M. A. & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessment: An update. *Journal of Research in Science Teaching, 33*(10), 1045-1063.

Shavelson, R. J., Solano-Flores, G. & Ruiz-Primo, M. A. (1997). Toward a science performance assessment technology. *Paper presented at the 7th EARLI Conference Athens, September*, 26-30.

Shepard, L. A. (1992). Commentary: What policy makers who mandate tests should know about the new psychology of intellectual ability and learning. In B.R.Gifford & M. C. O'Connor (Eds.), *Changing assessments. Alternative views of aptitude, achievement and instruction* (pp. 301-328). Boston: Kluwer Academic Publishers.

Silver, E. A. (1992). Assessment and mathematics education reform in the United States. *International Journal of Educational Research, 17*(5), 489-502.

Stebler, R., Reusser, K. & Pauli, C. (1994). Interaktive Lehr-Lern-Umgebungen: Didaktische Arrangements im Dienste des gründlichen Verstehens. In K. Reusser & M. Reusser-Weyeneth (Hrsg.), *Verstehen. Psychologischer Prozess und didaktische Aufgabe* (S. 227-259). Bern: Huber.

Stebler, R., Reusser, K. & Ramseier, E. (1997). Spitzenleistungen der Schweizer Siebtklässler im TIMSS-Experimentiertest. *Schweizer Lehrerinnen- und Lehrerzeitung SLZ*(10), 18-21.

Stebler, R., Reusser, K. & Ramseier, E. (1998). Praktische Anwendungsaufgaben zur integrierten Förderung formaler und materialer Kompetenzen - Erträge aus dem TIMSS-Experimentiertest. *Bildungsforschung und Bildungspraxis, 20*(1), 28-53.

Stenmark, J. K. (Ed.). (1991). *Mathematics assessment. Myths, models, good questions, and practical suggestions*. Reston, VA: The national council of teachers of mathematics.

Terwilliger, J. (1997). Semantics, psychometrics, and assessment reform: A close look at "authentic" assessment. *Educational Researcher, 26*(8), 24-27.

Vavrus, L. (1990). Put portfolios to the test. *Instructor, 100*(1), 48-53.

Wiggins, G. (1989a). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-713.

Wiggins, G. (1989b). Teaching to the (authentic) test. *Educational Leadership, 46*, 41-47.

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, 49*(8), 26-33.

Wolf, R. M. (1994). Performance assessment in IEA studies. *International Journal of Educational Research, 21*(3), 239-245.