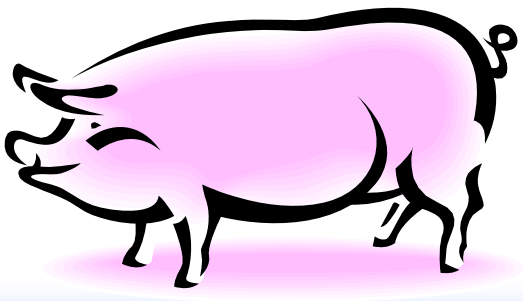
Vom Wiegen ….

Oder: Wie können Vergleichsarbeiten die pädagogische Praxis beeinflussen?

Eckhard Klieme

Deutsches Institut für Internationale Pädagogische Forschung

EMSE Tagung
Mainz, 6. Dezember 2007



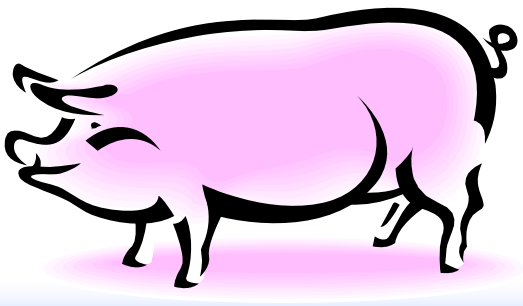
Bildungsmonitoring im Schulsystem

Ebene	Schüler	Klasse/Lehrkraft	Schule	Region	Land
	Diagnostik	**Evaluation**		**Bildungsbericht**	
Ziele	Förder-plan?	Curriculum?	Qualitätsrahmen, Schulprogramm	?	Strategische Ziele
Normativ	Lernstands-berichte	Dienstliche Beurteilung	Inspektions-bericht	?	Experten-kommission
Messung Standard. Urteile	Noten	-Skalen zur Unterrichtsbeobachtung - Befragungen aus mehreren Perspektiven		?	
Leistung-messung	Klassen-arbeiten, Tests ?	Parallelarbeiten	Orientierungs-/ Vergleichs-arbeiten		Standard-bezogene Vergleiche
Amtliche Statistik	Kerndatensatz				

dipf

These:

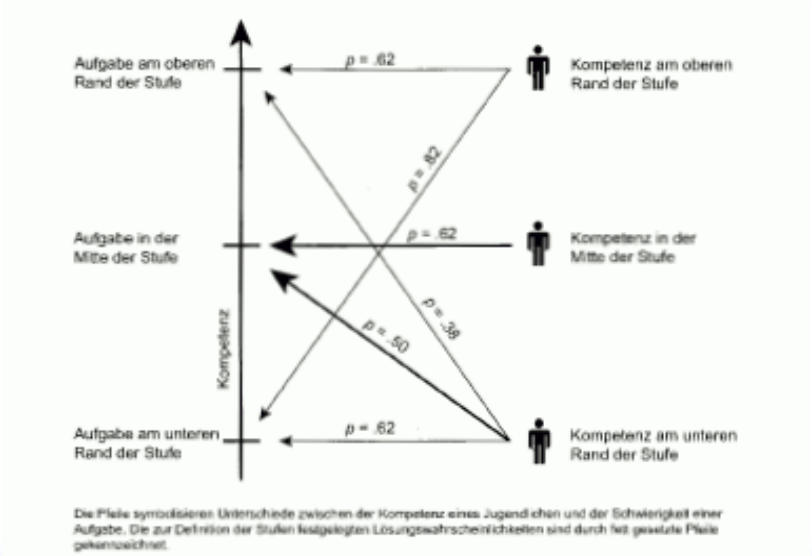
Bildungsmonitoring und Qualitätssicherung haben eine Hilfsfunktion.
Sie sind nicht die Therapie, sondern nur die Diagnose.



dipf

Gliederung

1. Gibt es das „Urmeter“ der Kompetenzmessung? Anmerkungen zum aktuellen PISA-Streit
2. Zwecke und Kontexte der Kompetenzmessung an Schulen: formatives vs. summatives Assesssment; Indikationen, Wirkungen, Risiken und Nebenwirkungen ….
3. Kompetenzmessung im Unterricht: E,pirische Erkenntnisse und das Beispiel des BEAR-Systems



Die Pfeile symbolisieren Unterschiede zwischen der Kompetenz eines Jugendlichen und der Schwierigkeit einer Aufgabe. Die zur Definition der Stufen festgelegten Lösungswahrscheinlichkeiten sind durch fett gesetzte Pfeile gekennzeichnet.

Abbildung 8.2: Lösungswahrscheinlichkeiten für eine Kompetenzstufe

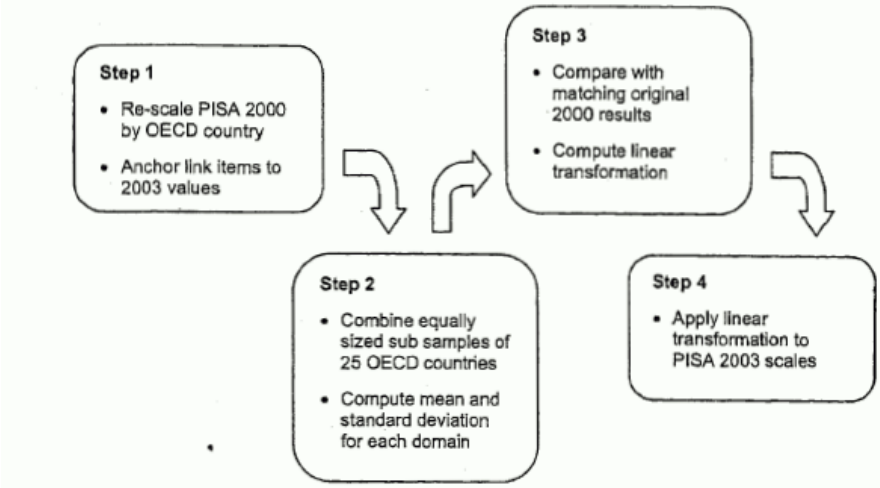


Figure 1. Steps involved in original linking of PISA 2000 and PISA 2003.

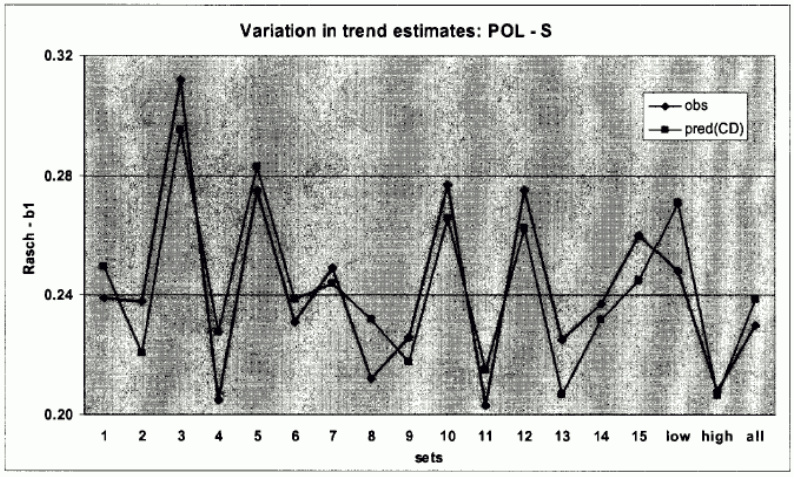


Figure 20. Trend estimates as estimated and predicted from the DIF measure (Science – POL)

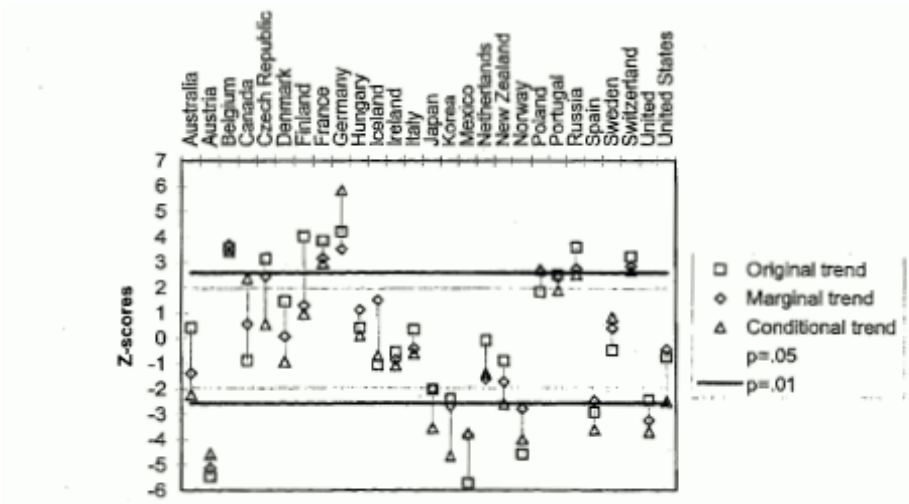


Figure 3. Three alternative trends between PISA 2000 and PISA 2003 in science by country

DFG-Projekt „Nutzung und Auswirkungen der Kompetenzmessung in mathematischen lehr-Lern-Prozessen

(Klieme/Rakoczy)

Hier:

Erkenntnisstand zu Auswirkungen von „high stakes tests“

- *Qualität der Tests:* Die eingesetzten Aufgaben erfordern häufig wenig produktive Leistungen und höhere Denkoperationen und entsprechen somit selbst den zum Teil eher oberflächlichen Standards der Teilstaaten nicht (Rothmann, Slattery, Vranek und Resnick 2002). Außerdem müssen die Tests ein so breites Spektrum des Curriculums abdecken, dass sie zu formativen Zwecken gar nicht genutzt werden können.

Klieme: Vom Wiegen .... EMSE-Tagung Mainz, 6.12.2007

- *Qualität der Tests:* Die eingesetzten Aufgaben erfordern häufig wenig produktive Leistungen und höhere Denkoperationen und entsprechen somit selbst den zum Teil eher oberflächlichen Standards der Teilstaaten nicht (Rothmann, Slattery, Vranek und Resnick 2002). Außerdem müssen die Tests ein so breites Spektrum des Curriculums abdecken, dass sie zu formativen Zwecken gar nicht genutzt werden können.

- *Qualität der Rückmeldungen:* Die Rückmeldungen, die Schulen, Lehrer und Schüler erhalten, sind häufig wenig informativ, rein normbezogen und eher kontrollierend. Damit widersprechen sie den Forderungen, die aus motivationspsychologischer Sicht an die Gestaltung von Rückmeldungen zu stellen sind (vgl. Abschnitt 2.1.2). Der

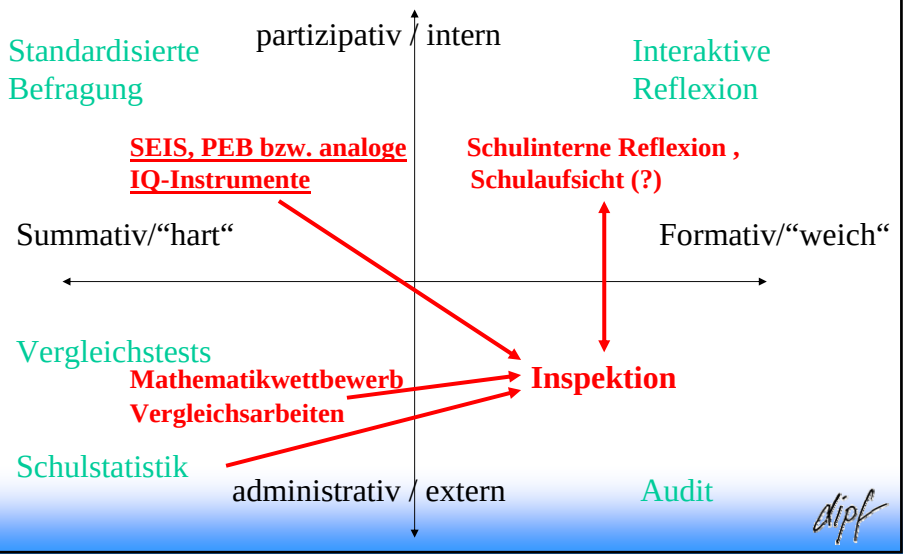
- *Qualität der Tests:* Die eingesetzten Aufgaben erfordern häufig wenig produktive Leistungen und höhere Denkoperationen und entsprechen somit selbst den zum Teil eher oberflächlichen Standards der Teilstaaten nicht (Rothmann, Slattery, Vranek und Resnick 2002). Außerdem müssen die Tests ein so breites Spektrum des Curriculums abdecken, dass sie zu formativen Zwecken gar nicht genutzt werden können.
- *Qualität der Rückmeldungen:* Die Rückmeldungen, die Schulen, Lehrer und Schüler erhalten, sind häufig wenig informativ, rein normbezogen und eher kontrollierend. Damit widersprechen sie den Forderungen, die aus motivationspsychologischer Sicht an die Gestaltung von Rückmeldungen zu stellen sind (vgl. Abschnitt 2.1.2). Der
- *Täuschungen:* Schulen versuchen gezielt, ihr Ergebnis zu beeinflussen – sei es durch einseitiges Testtraining, durch Exklusion schwacher Schüler oder durch Täuschungsmanöver. Dadurch wird nicht nur die Aussagekraft der Ergebnisse unterminiert, es zeigen sich auch negative Folgen für die Schulkultur.

Klieme: Vom Wiegen .... EMSE-Tagung Mainz, 6.12.2007

- *Qualität der Tests:* Die eingesetzten Aufgaben erfordern häufig wenig produktive Leistungen und höhere Denkoperationen und entsprechen somit selbst den zum Teil eher oberflächlichen Standards der Teilstaaten nicht (Rothmann, Slattery, Vranek und Resnick 2002). Außerdem müssen die Tests ein so breites Spektrum des Curriculums abdecken, dass sie zu formativen Zwecken gar nicht genutzt werden können.
- *Qualität der Rückmeldungen:* Die Rückmeldungen, die Schulen, Lehrer und Schüler erhalten, sind häufig wenig informativ, rein normbezogen und eher kontrollierend. Damit widersprechen sie den Forderungen, die aus motivationspsychologischer Sicht an die Gestaltung von Rückmeldungen zu stellen sind (vgl. Abschnitt 2.1.2). Der
- *Täuschungen:* Schulen versuchen gezielt, ihr Ergebnis zu beeinflussen – sei es durch einseitiges Testtraining, durch Exklusion schwacher Schüler oder durch Täuschungsmanöver. Dadurch wird nicht nur die Aussagekraft der Ergebnisse unterminiert, es zeigen sich auch negative Folgen für die Schulkultur.
- *Engführung der Lehr-Lern-Aktivitäten*: Die Ankündigung starker Konsequenzen hat zur Folge, dass Lehrkräfte ihren Unterricht in bedenklicher Weise an den Tests ausrichten. Dieser „washback-Effekt" (Cheng & Curtis 2004), der vielfach empirisch nachgewiesen wurde, betrifft die thematische Einengung, die Einengung der im Unterricht behandelnden Aufgaben und Anforderungen und die Reduktion der Lernzeit durch ausführliche Testvorbereitung. Auch Schülerinnen und Schüler schränken ihre Lernaktivität proaktiv auf Anforderungsbereiche der Leistungsmessung ein (Gielen, Dochy & Diereck 2003, S. 44).

- *Beeinträchtigung der Unterrichtsqualität*: Schon der Hinweis an Lehrkräfte, dass sie dafür verantwortlich gemacht würden, dass ihre Schülerinnen und Schüler bestimmte Standards erreichen, führte in einem Experiment von Deci et al. (1981) dazu, dass diese Lehrpersonen stärker kontrollierendes Verhalten zeigten, als Lehrkräfte, denen gesagt wurde, dass es keine Leistungsstandards gibt. Dies bedeutete, dass sie mehr sprachen, kritischer waren und den Schülerinnen und Schülern mehr Befehle gaben, während sie gleichzeitig weniger Wahlmöglichkeiten zuließen. Diese Verhaltensweisen wirken sich negativ auf die Motivation der Schüler aus (s. u.). In einer Stellungnahme zur aktuellen bildungspolitischen Diskussion kritisieren die Begründer der Selbstbestimmungstheorie daher die High-stakes-testings (Deci & Ryan 2002).

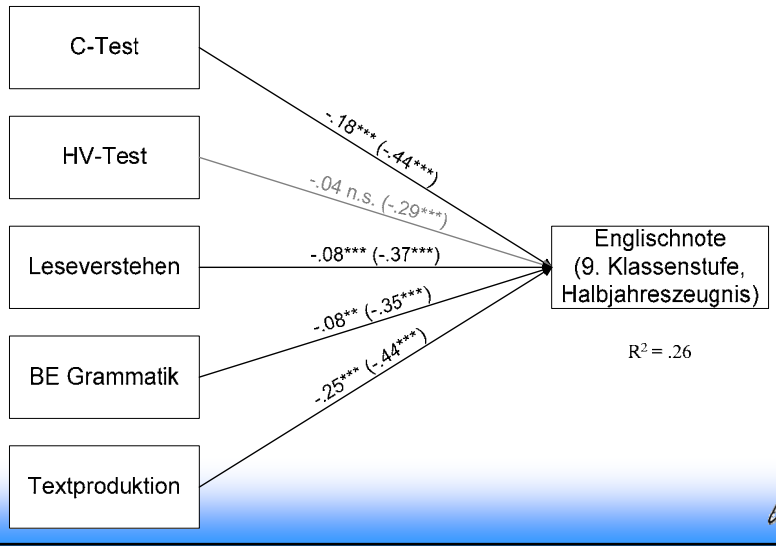
Typen der Schulevaluation



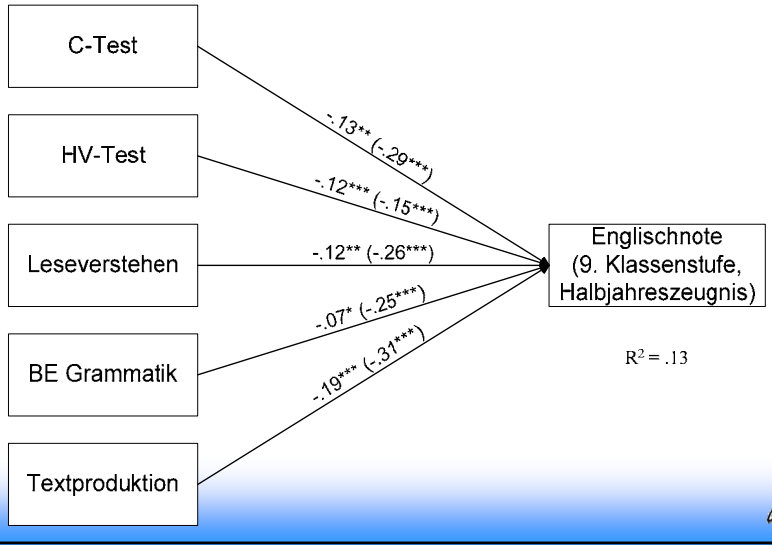
Unterrichtsrelevante formative Kompetenzmessung:

- Passung zwischen Unterricht und Bewertungskriterien
- Profile statt eindimensionaler Rankings
- Abbildung von Veränderungen
- Qualitative Diagnose des Verständnisses
- Einübung von Selbstregulation
- Feedback auch für „weiche“ Kompetenzen

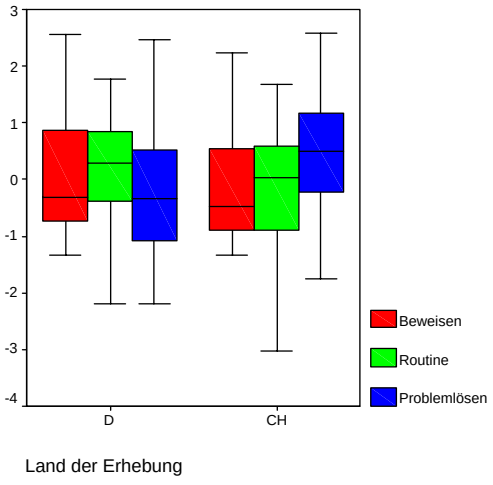
Problem 2: Didaktische Kulturen hier: Englischnote im Gymnasium



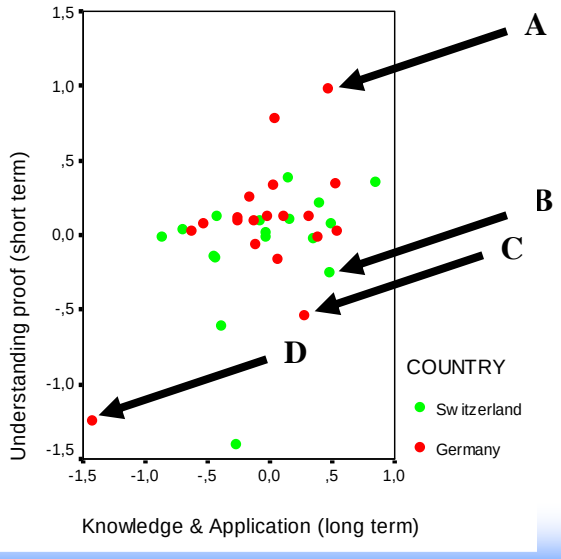
Didaktische Kulturen
Englischnote in der Realschule



Differenzierte Kompetenzprofile
(DFG-Projekt „Pythagoras“ (Klieme, Reusser et al.))



Profile der *Veränderung* von Kompetenzen



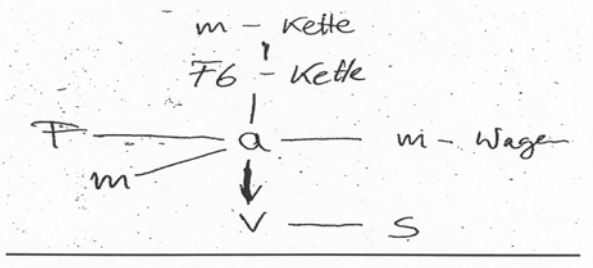
4. Jahrgangsstufe (IGLU)	9. Jahrgangsstufe (PISA)	Abschlussklassen der Sekundarstufe II (TIMSS)
	(V) Komplexes Modellieren und innermathemat. **Argumentieren** (1 %)	
(V) **Problemlösen** (7 %)	(IV) Umfangreiche Modellierungen auf der Basis anspruchsvoller begriffe (12 %)	(IV) Mathematisches **Argumentieren** (14 %)
(IV) Beherrschung der Grundrechenarten, räumliche Geometrie, begriffliche **Modellentwicklung** (35 %)	(III) **Modellieren** und begriffliches Verknüpfen auf dem Niveau der Sekundarstufe I (31 %)	(III) Mathematisches **Modellieren** und Verknüpfung von Operationen (34 %)
(III) Verfügbarkeit von Grundrechenarten und Arbeit mit einfachen Modellen (40 %)	(II) Elementare Modellierungen (32 %)	(II) Anwenden von einfachen **Routinen** (37 %)
(II) **Grundfertigkeiten** zum Zehnersystem, zur ebenen Geometrie und zu Größenvergleichen (17 %)	(I) **Rechnen** auf Grundschulniveau (17 %)	(I) Alltagsbezogene Schlußfolgerungen (15 %)
(I) Rudimentäres schulisches Anfangswissen (2 %)	Unter I (4 %)	

4. Jahrgangsstufe (IGLU)	9. Jahrgangsstufe (PISA)	Abschlussklassen der Sekundarstufe II (TIMSS)
	(V) Komplexes Modellieren und innermathemat. Argumentieren (1 %)	(IV) Mathematisches Argumentieren (14 %)
	(IV) Umfangreiche Modellierungen auf der Basis anspruchsvoller begriffe (12 %)	(III) Mathematisches Modellieren und Verknüpfung von Operationen (34 %)
	(III) Modellieren und begriffliches Verknüpfen auf dem Niveau der Sekundarstufe I (31 %)	Anwenden von einfachen Routinen (37 %) (II)
V) Problemlösen (7 %)	**(II) Elementare Modellierungen (32 %)**	**(I) Alltagsbezogene Schlußfolgerungen (15 %)**
(IV) Beherrschung der Grundrechenarten, räumliche Geometrie, begriffliche Modellentwicklung (35 %)	(I) Rechnen auf Grundschulniveau (17 %)	
(III) Verfügbarkeit von Grundrechenarten und Arbeit mit einfachen Modellen (40 %)	Unter I (4 %)	
(II) Grundfertigkeiten zum Zehnersystem, zur ebenen Geometrie und zu Größenvergleichen (17 %)		
(I) Rudimentäres schulisches Anfangswissen (2 %)		

Concept-Mapping

Beispiel für ein experimentbezogenes Schüler-Map

(DFG-Projekt Schecker, Klieme et al.)



Es geht um ein Experiment, bei dem ein Experimentierwagen auf einer Fahrbahn mittels einer angehängten Gliederkette beschleunigt wird. Die Kette wickelt sich langsam auf dem Boden auf, wodurch die Gewichtskraft der Kette ("FG-Kette") abnimmt. Die Beschleunigung "a" hängt von der Kraft und der zu beschleunigenden Masse ("m-Wagen"). Im Schüler-Map fehlt eine Relation zwischen dem Ort "s" und "FG-Kette" (und "m-Kette").

Computer – bezogene Fähigkeiten (Selbsteinschätzung)

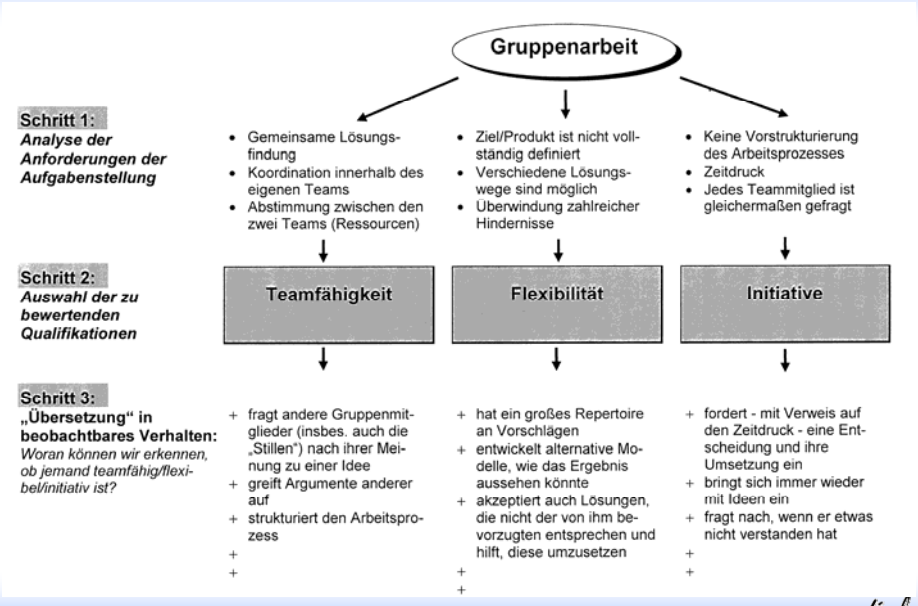
Wie gut bist du im Umgang mit dem Computer?

Wie gut bist du beim Schreiben eines Aufsatzes auf dem Computer?

Wie gut wärst du, wenn du eine Klassenarbeit am Computer schreiben müsstest?

Wenn du dich mit anderen 15-jährigen vergleichst, wie würdest du deine Fähigkeiten im Umgang mit dem Computer beurteilen?

Klieme: Vom Wiegen .... EMSE-Tagung Mainz, 6.12.2007



Unterrichtsrelevante formative Kompetenzmessung:

- Passung zwischen Unterricht und Bewertungskriterien
- Profile statt eindimensionaler Rankings
- Abbildung von Veränderungen
- Qualitative Diagnose des Verständnisses
- Einübung von Selbstregulation
- Feedback auch für „weiche“ Kompetenzen

Herausforderungen der Kompetenzmessung in Schulen:

1. Differenzierung zwischen Motivation/Interesse und kognitiver Leistung.
2. Differenzierung zwischen Leistungsdimensionen (z.B. Informationen ermitteln, reflektieren)
3. Differenzierung zwischen Niveaustufen:
   Was macht eine Aufgabe schwieriger?
   Was macht ein „besserer“ Schüler anders als ein „schlechterer“?
4. Realistische Zielsetzungen
5. Zuverlässige, breit angelegte Messung (vs. schnelle Information)
   → Unterscheidung zwischen unterrichtsbegleitender Diagnostik (z.B. Portfolios, Projektaufgaben) und standardisierten Tests (Vergleichsarbeiten)
6. Bewertung von „soft skills“ (z.B. durch Kopfnoten) ??

Das BEAR Assessment System
(Mark Wilson et al., Berkeley)

1. Developmental perspective
2. Match between Instruction and Assessment
3. Management by teacher
4. Quality evidence

Table 1
IEY Evidence and Tradeoffs Variable Scoring Guide

Score	Using Evidence	Using Evidence to Make Tradeoffs
	Response uses objective reason(s) based on relevant evidence to support choice.	Response recognizes multiple perspectives of issue and explains each perspective using objective reasons, supported by evidence, in order to make choice.
4	Response accomplishes Level 3 AND goes beyond in some significant way, such as questioning or justifying the source, validity, and/or quantity of evidence.	Response accomplishes Level 3 AND goes beyond in some significant way, such as suggesting additional evidence beyond the activity that would further influence choices in specific ways, OR questioning the source, validity, and/or quantity of evidence and explaining how it influences choice.
3	Response provides major objective reasons AND supports each with relevant and accurate evidence.	Response discusses *at least two* perspectives of issue AND provides objective reasons, supported by relevant and accurate evidence, for each perspective.
2	Response provides *some* objective reasons AND some supporting evidence, BUT at least one reason is missing and/or part of the evidence is incomplete.	Response states at least one perspective of issue AND provides some objective reasons using some relevant evidence, BUT reasons are incomplete and/or part of the evidence is missing; OR only one complete and accurate perspective has been provided.
1	Response provides only subjective reasons (opinions) for choice and/or uses inaccurate or irrelevant evidence from the activity.	Response states at least one perspective of issue BUT only provides subjective reasons and/or uses inaccurate or irrelevant evidence.
0	No response; illegible response; response offers no reasons AND no evidence to support choice made.	No response; illegible response; response lacks reasons AND offers no evidence to support decision made.
X	Student had no opportunity to respond.	

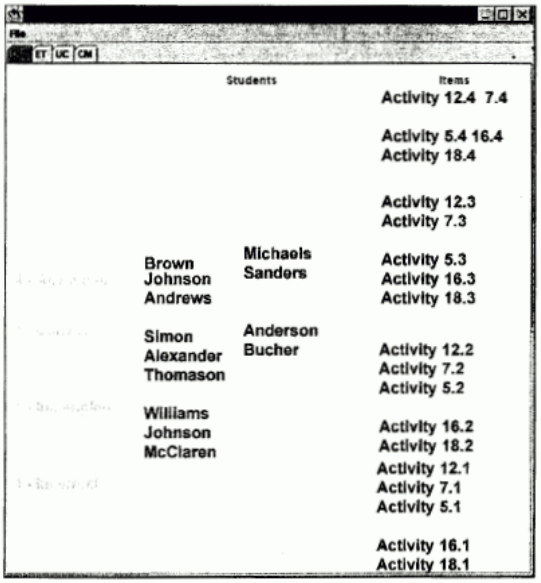


FIGURE 2
A progress map for a group of students' performances on the IEY Designing and Con-

144 LINKS BETWEEN LARGE-SCALE AND CLASSROOM ASSESSMENTS

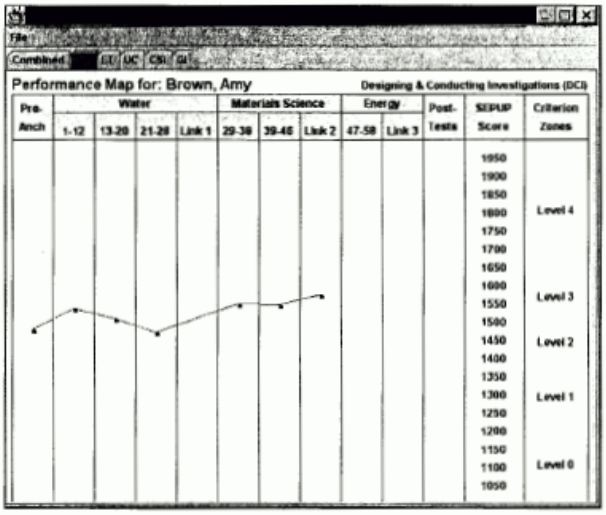


FIGURE 1

A progress map for an individual student's performance on the IEY Designing and Conducting Investigations variable.

Assessment is not an add-on to teaching and learning, it can be integral.

Richard Shavelson

Assessment is not an add-on to teaching and learning, it can be integral.

Richard Shavelson

Cross-walk between standards and assessment

Eva Baker