

Meine zentralen Botschaften vorweg

- Das **menschliche Urteil** ist das differenzierteste und sensibelste Instrument, das wir haben - Beobachtungen und Deutungen sind wegen ihrer Personabhängigkeit aber auch besonders fehleranfällig.
- Daraus wurde der Schluss gezogen, man müsse sie ersetzen durch **standardisierte Präzisionsinstrumente**. Das gelingt schon in der Medizin nicht, wo trotz aller High-Tech-Diagnose der Arzt immer noch die Deutungshoheit über die - trotz technischer Standardisierung interpretationsbedürftigen Daten - hat.
- Auch Leistungstests in der Pädagogik sind nicht besser als Thermometer usw. in der Medizin: Sie liefern Warnsignale, aber **keine Diagnosen** - es handelt sich um punktuelle Sondierungen.
- Deshalb müssen wir lernen, mit ihrer **Fehleranfälligkeit** zu leben - aber auch ihre Risiken zu reduzieren: durch **soziale Kontrolle** statt durch methodische Perfektionierung.

„Fieber zu messen ist noch keine Diagnose Fieber zu senken noch keine Therapie“

Zu den illusionären Versprechen standardisierter
Leistungsmessungen im Bildungswesen

Hauptvortrag von
Hans Brügelmann
(Universität Siegen)

zur Tagung der BWK
am 14.7.2011
an der IHK Siegen



Schule 2010 vs. 1960, 1980, 2000

Auffällig: die Allgegenwart von standardisierten Tests...

- bei der flächendeckenden **Sprachstandserhebung** mit Vierjährigen („Delfin-4“, „Pfiffikus-Haus“ in NRW)
- bei den landesweiten **Lernstandserhebungen** in 3. und 8. Klassen
- bei zentralen **Abschlussprüfungen** Ende Klasse 10 und 13
- im Rahmen **internationaler** Leistungsvergleiche zur Systemevaluation.

Standardisierung und Individualisierung

...sind widersprüchliche Anforderungen
der Bildungspolitik an Schule heute.

Der Konflikt kann von den Schulen nur aufgelöst werden,
indem Individualisierung auf eine methodische Differenzierung der
Lernwege beschränkt wird.

Individualisierung von **Inhalten** zur Entwicklung persönlicher
Potenziale wird erschwert, verliert an Bedeutung.

Output-Orientierung und Standardisierung

Wichtig:

- mehr Aufmerksamkeit dafür, was Schule bewirkt
- formalisierte Rechenschaftspflicht der Schule
- Misstrauen gegenüber subjektiven Urteilen

Problematisch:

- enge Fokussierung und Normierung der Ziele
- Missverständnis von Pädagogik als Technik
- Überschätzung standardisierter Methoden der Evaluation

Messbarkeit und Machbarkeit

...faszinieren nicht nur in der Psychologie und Pädagogik.

Verständlich: Neue Erkenntnisse der **Naturwissenschaften**
ermöglichen große Fortschritte in der Technik.

Aber technische Konzepte sind
auf den **Humanbereich** nicht einfach übertragbar.

Die Entwicklung der **Medizin** veranschaulicht das Dilemma -
und erst recht die Bewertung von **Kunstwerken** →



Das Zoom-Dilemma: Die Unschärferelation in den Sozialwissenschaften

Evaluation kann versuchen,

- additiv viele **Details** genau zu erfassen - oder
- ein strukturiertes **Gesamtbild** zu gewinnen.

Anders gesagt:

Wir stehen in der Spannung der konkurrierenden Anforderungen von

- **technischer Genauigkeit** einer Messung
- vs
- **inhaltlicher Bedeutung** ihres Ergebnisses

Textanalyse durch standardisierte Evaluation

A	4 x	B	1 x	C	1 x	D	1 x	E	10 x
F	2 x	G	1 x	H	3 x	I	5 x	J	-
K	1 x	L	6 x	M	1 x	N	7 x	O	-
P	3 x	Q	-	R	3 x	S	3 x	T	2 x
U	6 x	V	-	W	1 x	X	-	Y	-
Z	-								

Wissenschaftliche Analyse des Goethe-Gedichts „Über allen Wipfeln ist Ruh“ durch ein Institut für molekulare Poetik (Dürr, zit. nach Popp 2006, 11-12)

Das Lehrerurteil ist fehleranfällig - wie die Forschung zu Noten zeigt:

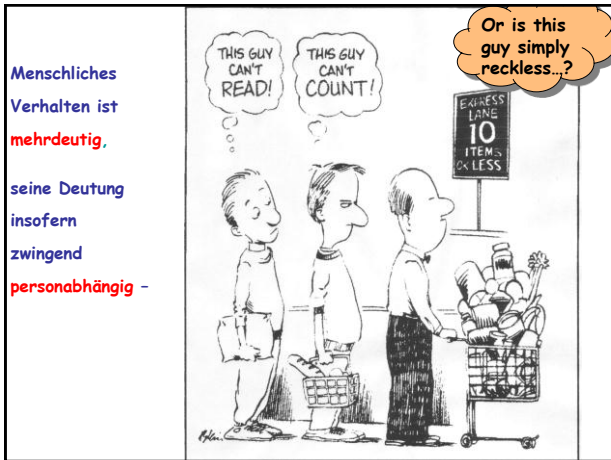
- **nicht valide**, d. h. nicht vorhersagekräftig bezogen auf zukünftigen Schul-, Ausbildungs- und Berufserfolg;
- **nicht verlässlich**, d. h. nicht stabil, sondern stark beeinflusst von äußeren Umständen (z.B. Reihenfolge);
- **nicht objektiv**, d. h. nicht unabhängig von den beurteilenden Personen (z.B. Sympathie);
- **nicht vergleichbar**, da sich die Bewertungen auf sehr unterschiedliche Maßstäbe und Schwellenwerte beziehen (z.B. Klasse im sozialen Brennpunkt vs. Villenviertel);

Standardisierte Tests statt persönlicher Urteile?

Vier Vorteile stecken als Potenzial in Tests:

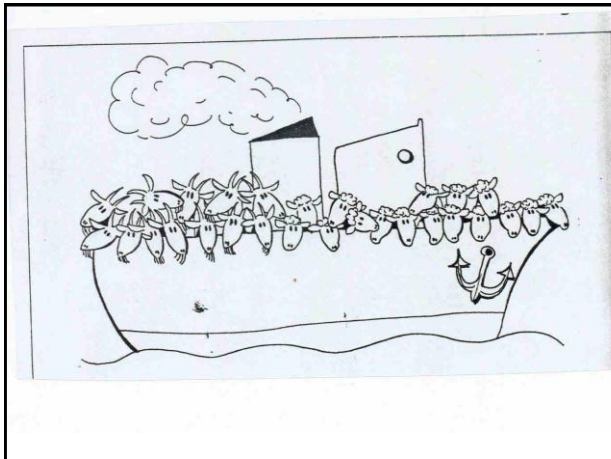
- **Fokussierung** der Datenerhebung
- **Transparenz** der Anforderungen
- **Kalibrierung** der Maßstäbe (Bezug auf Normstichproben)
- Unabhängigkeit von **persönlichen** Zufälligkeiten.

Insofern sind standardisierte Tests stärker als bisher in das Repertoire pädagogischer Leistungsbeurteilung einzubeziehen.



Tests können deshalb das Lehrerurteil nicht ersetzen, denn...

- sie müssen sich auf bestimmte **Aufgabentypen** und **Auswertungsformen** beschränken („Standardisierung“);
- sie konzentrieren sich inhaltlich auf **wenige Ausschnitte** der Leistungsprofils (Messbarkeit; Ökonomie);
- ihre Ergebnisse sind bei **punktuellem Erhebung** ebenfalls nicht verlässlich;
- ihre **prognostische Validität** ist ähnlich unsicher wie die der Noten.



Die Frage lautete:

Wie alt ist der Kapitän?

Was hast du dir überlegt? Wie bist du auf deine Antwort gekommen?

ich habe die Fläche gezählt weil sonst käme man nie auf das ergebnis.

Was hältst du von dieser Aufgabe? Wie gerne hast du sie bearbeitet?

ich finde sie ist spannend.

Was hast du dir überlegt? Wie bist du auf deine Antwort gekommen?

Ich habe alle Schaffe und alle Ziegen zusammen gezählt. Und dan bekam ich die Antwort.

Was hältst du von dieser Aufgabe? Wie gerne hast du sie bearbeitet?

Ja ich fürchte dass dass eine Scherzfrage ist aber vielleicht auch nicht

So zu sagen wie auch alle anderen Aufgaben die wir in der Schule machen.

Antwort: 20 Jahre alt

Was hast du dir überlegt? Wie bist du auf deine Antwort gekommen?
Weil ein Schaf nicht viel älter
wesen kann.

Was hältst du von dieser Aufgabe? Wie gerne hast du sie bearbeitet?
Ich finde die Aufgabe lustig und
ein bisschen komisch.

Antwort: Der Kapitän ist 28 Jahre alt

Was hast du dir überlegt? Wie bist du auf deine Antwort gekommen?
Weil wenn man Geburtstag hat schenkt man
30 Rosen oder eben hat 12 Ziegen und 16
Schafe. Dann habe ich es zusammen gezählt
und ich habe beschlossen das der Kapitän
28 Jahre alt ist. PS: Alles Gute

Was hältst du von dieser Aufgabe? Wie gerne hast du sie bearbeitet?
Die Aufgabe ist mir unbekannt gewesen
aber man muss nur überlegen.
Ich habe die Aufgabe gerne gemacht
weil man musste überlegen.



Auch die Schulkultur prägt

Alter	gerechnet haben
Kiga und 1. Klasse	11 %
2. Klasse	32 %
3. Klasse	54 %
4. Klasse	58 %
5. Klasse	46 %

Was können wir daraus lernen?

Jeder rekonstruiert Aufgaben auf seine Weise

- anders als andere Lerner
- anders als die Autoren der Aufgaben
- nicht fassbar in eindeutigen Bewertungen.

Das Grundproblem ist die Mehrdeutigkeit menschlichen Verhaltens, das in der Standardisierung von Aufgabenstellung und -auswertung nur überdeckt wird.

- Was bedeutet das für Schulen & Betriebe?**
1. Behutsamkeit im Umgang mit **Noten** und **Tests** - selbst wenn nur als Grobfilter genutzt.
 2. Selbstkritischer Umgang mit **eigenen Bewertungen** - selbst wenn man viel Erfahrung hat.
 3. Ersatz/ Ergänzung von Noten durch differenzierte **Berichte**, die Subjektivität transparent machen.
 4. Minimierung des unvermeidlichen **Fehlerrisikos**, der Wahrscheinlichkeit von Fehlurteilen, durch →

Vier schlichte Vorkehrungen:

- Kombination verschiedener Aufgaben(typen)
 - Stärkung der **Validität**, der inhaltlichen Gültigkeit
- Nutzung von Daten aus unterschiedlichen Situationen
 - Steigerung der Reliabilität, der **Zuverlässigkeit**
- Einbeziehung **mehrerer Perspektiven** bei der Bewertung
 - Reduktion von subjektiver Willkür
- **dialogische** Rahmung statt Urteil „von oben“
 - Sicherung von Fairness

Fazit

Menschen können nur von Menschen beurteilt werden.

Produktkontrolle, TÜV und Stiftung Warentest sind keine geeigneten Modelle für pädagogische Evaluation.

Der Preis für die Sensibilität des menschlichen Urteils ist seine Subjektivität und Fehleranfälligkeit.

Wir können diese nicht verhindern -
wohl aber das Risiko verringern,
dass ihre Auswirkungen den Betroffenen schaden:
den KandidatInnen wie den auswählenden Einrichtungen.

