Grundsätzliches zur doppelten Funktion der schulischen Leistungsmessung

Wer an einem Seminar zur Leistungsmessung teilnimmt, war nicht nur erfolgreich (sonst säße er ja schließlich nicht im Seminar), sondern hat auch erfolgreich gelernt, daß die gängigen Verfahren der Leistungsmessung (Klassenarbeiten, Klausuren, Referate und Hausarbeiten) nicht-erfolgreiche Lerner eliminieren. Einfach so.
Diese erfolgreiche Anpassung an die letztendlich einzig wesentliche pädagogische Spielregel ist zwar akademisch überlebenswichtig, sie beruht aber nichtsdestoweniger auf einem arg naiven Kinderglauben à la „Die Schule/Uni hat immer recht", „Der Lehrer hat immer recht", „Wenn ich eine schlechte Note bekomme, geschieht mir nur recht".
Spätestens im Referendariat machen einem die Mentoren und Seminarleiter freundlich aber bestimmt deutlich, daß der Umgang mit der Macht der Note wohl gelernt sein will. Auch das geht wieder über Anpassung: Wenn zu viele Arbeiten unterm Strich sind, war die Arbeit unterm Strich, d.h. sie muß wiederholt werden. Und das wirft nicht nur ein schlechtes Licht auf den Lehrer, sondern bedeutet auch Arbeit. Also lernt man als ReferendarIn von gestandenen Pädagogen aus der Praxis für die Praxis, wie man Klassenarbeiten so konzipiert und auswertet, daß die Frage nach Qualität (und Legitimation) möglichst gar nicht erst gestellt wird. — Das würde ja auch nur die zweifellos notwendige Anpassung der Schüler an die Selektionspraxis behindern —.
Dieses unverdrossene Streben nach Einhaltung von konsensuellen Qualitätsstandards in der Leistungsmessung hat zu einer breiten Palette von zwei erprobten Verfahren zur Qualitätssicherung geführt: der Notenspiegel wird einfach ausgezählt, beim Fehlerquotienten kommen schon die höheren arithmetischen Operationen der Multiplikation und Division in dreisatz-komplexer Ordnung zur Anwendung. Weitergehende statistische Analysen würden die Lerner nur verunsichern, eventuell sogar aufmüpfig machen. Da verzichtet man doch besser gleich drauf.
Solange Auswertungen sehr zeitaufwendige Berechnungen mit dem Taschenrechner erforderten, konnten detailliertere und damit aufwendigere Analysen billigerweise von keinem praktizierenden Lehrer mit eh schon sehr hoher Stunden- und Korrekturbelastung erwartet werden. Seit die entsprechenden Berechnungen jedoch in Standard-Software-Paketen enthalten sind — und überdies auch noch mühelos grafisch dargestellt werden können — gibt es eigentlich keinen vernünftigen Grund mehr, auf die demokratischere Transparenz zu verzichten, die statistische Prozeduren nun einmal anschaulich machen.
Da erfahrungsgemäß bei sprach- und literaturwissenschaftlich Orientierten das mathematische Interesse doch eher gedämpft ist, sollte der erste Schritt in Richtung seriöse Datenauswertung über die Visualisierung der Basis-Tabellen gehen. Da alle gängigen Tabellenkalkulationen eine grafische Komponente haben, kann man sich mit wenig mehr als einem Mausklick einen ersten — und manchmal sogar auch ausreichenden — Überblick über Verteilung und Zusammenhänge der Daten verschaffen.
Wichtig dabei ist die doppelte Fragestellung nicht mehr nur nach dem Abschneiden der einzelnen Testanden, sondern auch nach der Meßgenauigkeit und Aussagekraft der einzelnen Items. Die der EDV inhärente Auswertungsökonomie läßt es unsinnig erscheinen, auf diese für den zukünftigen Unterricht überaus wichtigen Informationen zu verzichten. Die ohne großen Aufwand mögliche Item-Analyse in grafischer Darstellung gibt dem Testautor wichtige Hinweise über offensichtliche Schwächen bei der Aufgaben- und Item-Konstruktion. Die entsprechende Grafik macht unmittelbar sinnfällig, ob die vorliegende Verteilung der Datenwerte annähernd normalverteilt ist oder nicht (ist die Verteilung ein- oder mehrgipflig, ist sie gestaucht oder gestreckt, ist sie vielleicht linksschief, wie stark ist die Streuung …). Die Grafik beantwortet auch ohne theoretische Zusatzabstraktion Fragen nach einem etwaigen Positionseffekt oder nach der inneren Konsistenz der Items (Reliabilität) und natürlich auch nach dem Differenzierungspotenzial der Items (Trennschärfe).
Wichtiger als all diese technischen Detailinformationen ist jedoch der wissenschaftliche Aufklärungswert des für die statistische Auswertung verwendeten stochastischen Modells: Es macht schon einen nicht nur atmosphärischen Unterschied, ob ein deterministisch denkenderTestautor

einem Testanden sagt „Du hast 17 Fehler, also bist du 5“ oder ob ein mit Wahrscheinlichkeiten operierender Testautor einen Testanden darüber informiert, daß er mit 17 Fehlern zur schwächsten 7 %-Gruppe zählt.
Zugegeben, die doppelte Analyse über die *Items* ***und*** über die *Vpn* ist anfangs vielleicht etwas gewöhnungsbedürftig, aber die hiermit erreichte demokratische Transparenz der Notengebung hat einen mächtigen solidarisierenden Effekt, weil aus dem distanzierten Lehrer mit absolut gesetzter (Pseudo-)Autorität der ebenfalls noch lernende Partner in der pädagogischen Interaktion wird, der als Fachmann für das Lern/Lehrgeschäft mit den Lernenden zusammen versucht, die Lern- und Testleistungen zu optimieren.

D I E E M P I R I S C H E P A R A D O X I E

Theorien werden konstruiert, um die Wirklichkeit zu verstehen. Die Wirklichkeit wird vermessen, um die Theorien zu verbessern. Diese Interaktion verweist auf das konstitutive Dilemma jeder Wissenschaft: Ohne theoretische Vor-Annahme kann ich die Wirklichkeit nicht kategorisieren und strukturieren, ohne Meßdaten kann ich die theoretischen Vorstellungen nicht an ihren Geltungsbereich im gewählten Wirklichkeitsausschnitt optimieren. Dieses Verhältnis zwischen Theorie und Wirklichkeit beinhaltet noch keinen Widerspruch, sondern läßt sich als Methode hinreichend genau beschreiben mit Begriffen/Verfahren der Iteration und sukzessiven Approximation.
Die Empirische Paradoxie kommt erst dann ins Spiel, wenn ich stochastische (wahrscheinlichkeitstheoretische) Modelle hypostasiere (als gültig festsetze) und die empirisch festgestellten Verteilungen daraufhin überprüfe, ob sie diesen theoretischen Erwartungen entsprechen bzw. mit welcher angebbaren Wahrscheinlichkeit sie davon abweichen. Dabei liegt der Widerspruch nicht in den theoretischen vs. den beobachteten Verteilungen, sondern vielmehr darin, daß die gängigen statistischen Verfahren üblicherweise entweder die Datenlieferanten (Vpn) oder die Reaktionen (Testitems) festschreiben, um die jeweilige Kehrseite der Medaille zu überprüfen, obwohl doch der gesunde Menschenverstand weiß, daß verfahrenstechnisch die untersuchten Personen durch ihre Meßwerte definiert werden und umgekehrt.
Bei unserem Präpositions-Test bedeutet dieses Paradoxon in positiver Umkehrung, daß wir die Item-Ergebnisse behandeln, als kämen sie von normalverteilten Vpn, und die Vpn, als lieferten sie normalverteilte Item-Ergebnisse. Wir bewerten also die Vpn so, als wären die Item-Ergebnisse zuverlässig (reliabel) und gültig (valide) — obwohl doch die Diagramme erhebliche Zweifel an der Qualität der Tests aufkommen lassen. Andererseits können wir die Qualität der Tests, d.h. der Items, nur abschätzen, indem wir eine irgendwie angenäherte Normalverteilung der Vpn annehmen, was ihre Beherrschung der englischen Präpositionen angeht.
Zwar haben mathematische Statistiker dieses Problem seit langen erkannt und auch gelöst, aber die erforderliche „doppelte Buchführung“ ist noch nicht durchgeschlagen auf die statistischen Standard-Programm-Pakete. Da (prospektive) FremdsprachenlehrerInnen in aller Regel eher lernzielorientierte Klassenarbeiten verwenden und nicht etwa an der Eichung standardisierter Tests beteiligt sind, müssen wir uns mit den handelsüblichen Verfahren behelfen und die gleichen Meßdaten einmal über die Items und einmal über die Vpn analysieren. Da bei den üblichen Lerngruppengrößen und Testumfängen eh keine mathematisch saubere Verteilung anzunehmen ist, müssen wir bei der Auswertung damit leben, daß die finale Notenzuweisung doppelt angenähert ist. Die Verantwortung für die Notengebung bleibt uns voll erhalten. Wir können nur für eine objektive Basis sorgen und ihre Zuverlässigkeit transparent machen.
Der wohl wichtigste Effekt der routinemäßig eingesetzten Verfahren zur Überprüfung der empirischen Verteilungen, der Positionseffekte, der Inter-Item- und Inter-Test-Korrelationen sowie der überschlägigen Trennschärfe liegt wohl in der Selbstbescheidung des Test-Autors und –Auswerters. Nach all diesen Berechnungen fällt es ganz einfach schwer zu glauben, daß das

Testergebnis des einzelnen Testanden seine Leistungsfähigkeit ungebrochen von Qualitätsmerkmalen des Tests abbildet.
Die Orientierung am Klassenspiegel und am Fehlerquotienten führt fast immer zu einer ausreichend genauen Differenzierung der Testleistungen. Diese Fixierung auf das Abschneiden der Testanden verstellt aber allzusehr den Blick für die bei hausgemachten Tests oft überraschend ausgeprägte und im ersten Anlauf auch kaum zu erklärende Unzuverlässigkeit vieler Test-Items und damit des Tests überhaupt.
Diese völlig normale Unzuverlässigkeit von selbst erstellten Tests läßt sich einfach nicht vermeiden. Die hohe Zuverlässigkeit von standardisierten Intelligenz- und Eignungstests mit Reliabilitätskoeffizienten über r = 0,9 ist das Ergebnis einer durch viele Probeläufe gegangenen Validierung. Im lernzielorientierten Testbereich stehen dafür weder die Zeit noch die Mittel zur Verfügung. Um auf Anhieb einen Reliabilitätskoeffizienten von r = 0,7 (der etwa die Hälfte der Streuung der Werte aufklärt) zu erzielen, muß man schon ein guter Kenner der Testinhalte und –verfahren sein und auch den Leistungsstand der untersuchten Personengruppe sehr gut einschätzen können. Die mangelnde Aufklärung von formalen Verfahren als Argument für ihre Abschaffung zu verwenden hieße jedoch, das Kind mit dem Bade auszuschütten. Der Rückfall in subjektive Willkür läßt sich auch durch paternalistische Ignoranz nicht entschuldigen: Die gängigen Testverfahren sind zwar unzuverlässig, aber sie sind das Zuverlässigste, was wir haben. Und darüber kann man nicht streiten.

Hier in Rezeptform die Auswertungsschritte für eine faire Bewertung:

1. ***Tabellen*** von Teil-Tests mit Zusammenfassungen über Vpn und Items (Summe, AM, SD, % jeweils in Spalten und Zeilen);
2. ***Sortieren*** der Meßwerte nach Vpn und Items. Damit Item-Analyse-Grafiken erstellen über Fehlerhäufigkeiten, Positionseffekte und globale Trennschärfe;
3. Berechnen der ***Reliabilität***skoeffizienten;
4. Die ***Korrelationen*** zwischen Test-Teilen in Streudiagrammen visualisieren;
5. Eventuell Mittelwertsunterschiede mit dem ***„t-Test“*** auf Signifikanz überprüfen;
6. Die ***Fehlerbandbreite*** (ohne statistische Ausreißer) in 5 bzw. 6 gleichgroße ***Noten***bereiche einteilen.

- ***…und in Zweifelsfällen noch einmal den Spruch bedenken: Ein Test mißt, was seine Items messen. Das ist zuallererst die Kompentenz des Testautors — und in zweiter Linie auch die Kompetenz der Testanden.***