Messen und Zensieren im Sportunterricht

Meinhart Volkamer

Verlag Karl Hofmann 7060 Schorndorf

2. Der Unterschied von „Messen" und „Bewerten"

Die Voraussetzung für jede *Bewertung* (im Sinne von „*Zensur*") ist die möglichst objektive und vor allem quantitative Erfassung der betreffenden Leistung. Dieses quantitative Erfassen bezeichnen wir als „messen". Messen bedeutet, „den Objekten oder Ereignissen — bestimmten Regeln gemäß — entsprechende Zahlen zuzuordnen (CLAUS / EBNER, 21). Durch diese Zahl wird dem gemessenen Gegenstand oder Ereignis ein ganz bestimmter und unverwechselbarer Platz in dem vorher definierten Zahlenraum zugewiesen; wer die Zuordnungsregel kennt, der kann ein bestimmtes Objekt oder Ereignis im entsprechenden Zahlenraum eindeutig lokalisieren.

Wir müssen uns klar machen, daß zwischen *Messen* und *Bewerten* (im Sinne von Zensieren) ein wesentlicher Unterschied besteht. Machen wir uns das an folgendem Beispiel deutlich:

Wir erfahren:

„Schüler A ist 5 m weit gesprungen". „5 m" ist ein Meßwert, der sich definitorisch aus den Regeln des Weitsprungs ergibt und mit einem bestimmten Meßsystem — nämlich dem Metermaß — erfaßt wird. Die Regeln des Weitsprungs (was wird gemessen) und die Eigenart des Metermaßes weisen dem konkreten Ereignis „Weitsprung" einen bestimmten und unverwechselbaren Platz in dem Zahlenraum, der durch „Meter" gekennzeichnet ist, zu. Der Meßwert „5 m" beschreibt also das Ergebnis der Handlung „Weitsprung" nach einer vorher vereinbarten Weise. Die Beschreibung des Ergebnisses durch die (in m, cm ausgedrückte) gesprungene Weite ist aufgrund von Vereinbarungen typisch für diese Sportdisziplin — es wäre unsinnig, die Leistung des Weitspringers mit der Stoppuhr erfassen zu wollen.

D. h. die Leistungsmessung ist *handlungsspezifisch* und macht keine Aussage über die Person, die diese Leistung erbracht hat. Die Aussage „der Sprung war 5 m weit" ist sinnvoll auch *ohne* jede Information über denjenigen, der gesprungen ist.

Die sportlichen Leistungen werden — handlungsspezifisch und aufgrund vorhergegangener Vereinbarungen — nach Meter, Sekunde, Kilogramm (in den sog. C-G-S-Sportarten), in den Spielen durch das Zählen von Toren (etwa im Fußball) oder von Fehlerpunkten (z. B. im Tennis) oder durch das Zuteilen von Haltungspunkten (z. B. im Kunstturnen) etc. gemessen.

D. h. auch das Zählen von Toren ist ein Meßvorgang: das Zählen ist eine der einfachsten Arten, einem Ergebnis einen Platz in einem bestimmten Zahlenraum zuzuweisen.

Ebenso ist die „Wertung" einer Turnübung in diesem Sinne eine spezifische Art des Messens (der quantitativen Leistungsfeststellung), die sich vom Messen eines Weitsprungs nur durch die höhere Komplexität der Handlung, durch geringere

Operationalisierbarkeit des Ereignisses (Haltung, Rhythmus, Flüssigkeit etc.) und durch geringere Objektivität des Meßinstrumentes (Schätzung durch Kampfrichter) unterscheidet (Man versucht, die Genauigkeit des „Meßinstruments" z. B. durch intensive Schulung der Punktrichter zu erhöhen; trotzdem treten immer noch — auch bei internationalen Wettkämpfen — erhebliche Unterschiede zwischen verschiedenen Kampfrichtern auf, weil es kein personunabhängiges Meßinstrument, wie z. B. das Metermaß, für die Feststellung einer turnerischen Leistung gibt).

In der Vergabe von Wertungspunkten beim Turnen wird der gezeigten Leistung ein bestimmter Zahlenwert zugeordnet mit dem Anspruch, daß die zugeordneten Zahlenwerte diese Leistung in einem definierten Zahlenraum eindeutig lokalisieren, mit der Folge, daß verschiedene Leistungen miteinander verglichen werden können.

Wenn wir also in den Sparten wie Turnen, Gymnastik, Kunstspringen etc. von „bewerten" sprechen, so hat das Wort die Bedeutung von „messen" und unterscheidet sich von dem, was wir unter „bewerten = zensieren" verstehen. Dieser Bedeutungsunterschied wird häufig nicht gesehen, wodurch die ohnehin verworrene Diskussion um die Zensur noch verworrener wird. „Die Trennung von Leistungsmessung und Leistungsbewertung ist Hauptproblem jeder Prüfungstheorie" (Füller 1975, 26) (Genauer spricht man beim Turnen auch nicht von einer Bewertung sondern von einer Wertung).

Wenn im folgenden der Begriff „Bewertung" benutzt wird, so ist damit das Erteilen von Zensuren gemeint. Im Gegensatz zur Messung, die die Höhe einer Leistung quantifizieren, in einen Zahlenraum abbilden will, wird mit der Zensur eine Bewertung der Leistung — „sehr gut, gut, ..., ungenügend" — ausgesprochen. Diese Bewertung wird zwar auch in einer Zahl verkürzt zusammengefaßt („sehr gut" = 1, etc.), aber die Ziffernzensur „1" hat einen grundsätzlich anderen Aussagewert als etwa „4,50 m im Weitsprung". Während die Messung eines Weitsprungs oder die Wertung einer Turnübung eine *Quantifizierung* darstellt, ist die Zensierung ein Akt der *Qualifizierung*.

Um den Unterschied zwischen Messen und Bewerten (= Zensieren) zu verdeutlichen, nehmen wir wieder unser Beispiel auf: „Schüler A springt 5 m weit." Es könnte uns nun jemand fragen: „Wie beurteilen sie diesen Sprung?" Welche Zensur würden Sie dafür geben? Ist das eine gute oder nur eine mäßige oder gar schwache Leistung?"

Um diese Fragen beantworten zu können, müßten wir individuelle Daten des Schülers (z. B. Alter, Geschlecht, körperliche Voraussetzungen ...) kennen. 5 m ist für ein Mädchen schwerer zu erreichen als für einen Jungen, für einen Großen leichter als für einen Kleinen etc. Das allein genügt aber noch nicht: wir müssen auch noch wissen, was Schüler mit vergleichbaren Voraussetzungen durchschnittlich in dieser Disziplin leisten. Erst wenn wir wissen, daß A 14 Jahre alt ist, und 14jährige Jungen im Durchschnitt 4 m weit springen, können wir sagen, daß der Sprung von 5 m überdurchschnittlich gut ist, während die 5 m für einen 18jährigen allenfalls dem Durchschnitt entsprechen. D. h. aufgrund eines Vergleichs

mit den Leistungen einer Bezugsgruppe wird festgestellt, ob 4,50 m im Weitsprung „sehr gut" (z. B. für einen 12jährigen), „befriedigend" (z. B. für einen Abiturienten im Leistungsfach Sport) oder „mangelhaft" (z. B. für einen Sportstudenten) sind. „Die schulische Leistungsbeurteilung erfolgt (nämlich) in einem sozialen Raum" (Potthoff 1974, 20).

Damit hat die Zensur auch notgedrungen einen anderen Aussagewert als eine Leistungsfeststellung: Eine „2" besagt: „Im Vergleich zu anderen ist deine Leistung über dem Durchschnitt; es gibt noch bessere als dich, aber es gibt auch viele, die schlechter sind." Die Leistung, auf der dieses Urteil beruht (z. B. 4,50 m im Weitsprung), kommt in der Zensur nicht mehr zum Ausdruck.

In der Zensur wird die konkrete individuelle Leistung des Schülers in einem *sozialen* Raum lokalisiert, die Zensierung stellt immer einen *Vergleich mit anderen* dar, sie ist eine sozial gebundene Aussage. Die Information „Schüler A ist ‚gut' im Weitsprung" sagt nichts über das tatsächliche Handlungsergebnis aus, sagt uns nicht, wie weit er gesprungen ist; wir können dem Urteil nur entnehmen, daß A besser weitspringt als die meisten anderen vergleichbaren Schüler.

Während die Leistungs*messung* (= Quantifizierung) *handlungsspezifisch* (m, sec, kg, Tore etc.) erfolgt, ist die *Zensierung* (= Qualifizierung) *handlungsunspezifisch*: Die Zensurenskala ist auf jedes Handlungsgebiet anwendbar; ein Weitsprung kann ebenso „gut" sein wie ein Deutschaufsatz, wie Vokabelkenntnisse oder die Lösung einer Rechenaufgabe.

Eine Leistungsmessung bietet eine *sachgebundene Information* (z. B. 6 m weit), eine Zensur ist eine *personengebundene Aussage* (A ist im Vergleich zu ... gut.).

Das Zensieren = Bewerten unterscheidet sich noch durch ein weiteres Merkmal vom Messen: Der Meßwert ergibt sich in vielen Fällen handlungsspezifisch *unmittelbar* aus der Handlung selbst (der Springer hat die Höhe übersprungen, das Bandmaß zeigt 20 m an, der Ball hat die Torlinie überschritten, etc.); die Zensur dagegen wird von einem Dritten, einer Beurteilungsinstanz *zugeteilt*, wie es typisch in unseren Sprachformeln zum Ausdruck kommt: „Für diese Leistung bekommt der Schüler ..." oder „Dafür gebe ich ihm ...", im Gegensatz zu: „Seine Weitsprungleistung ist ...".

D. h. Bewertung ist im doppelten Sinne ein sozial eingebundener Vorgang,

a) dadurch, daß die individuelle Leistung mit der Leistung anderer verglichen wird,
b) dadurch, daß sie dem Zensierten ausdrücklich von einer anderen Person zugeteilt wird.

Die Zensur informiert die Adressaten (Schüler, Eltern, Behörde etc.) nicht über die objektive Leistung des Schülers, sondern über seine relative Leistung im Vergleich zu anderen.

Diese Unterscheidung von Leistungsmessung und Leistungsbewertung ist zur Vermeidung von Mißverständnissen äußerst wichtig und erscheint fast trivial, sie wird aber auch in der Fachliteratur häufig nicht sauber durchgehalten, was z. T. zu grotesken Mißverständnissen führt:

So stellt SCHRÖTER (1977) fest: „Problematisch bleibt natürlich die Notengebung. Leistungsmessung und Notengebung sind nicht identisch" (S. 94). Richtig, nur hält er selbst diese Unterscheidung nicht durch; das wird in einem späteren Kapitel mit der Überschrift „Zensuren gibt es nicht nur in der *Schule*" deutlich: Als Beleg dafür, daß wir „auch außerhalb von ihr mit Zensuren aller Art" umstellt sind, daß unser ganzes Alltagsleben mit ihnen durchsetzt ist, führt er u. a. auch Beispiele aus dem Bereich des Sports an:

„Bei jedem Fußballspiel werden klare, eindeutige Ziffernzensuren erteilt (Torverhältnis), die in ihrer Gesamtheit (Tabelle) auch etwas über den ligainternen Stand aussagen" (S. 173). Beim Fußballspiel wird gezählt, wie oft der Ball ins Tor geschossen wird. Diese zahlenmäßige Feststellung erfolgreich abgeschlossener Angriffe wird hier gleichgesetzt mit der Erteilung von Zensuren, mit der Beurteilung (durch wen?), ob eine Handlung gut oder vielleicht nur befriedigend war. Aus dem Vollzug des Spiels ergibt sich im Torereignis eine Leistungsfeststellung, die in keiner Weise in ein Beurteilungssystem transformiert werden muß. Erst der Lehrer in der Schule, den es weniger interessiert, ob ein Schüler ins Tor getroffen hat, als vielmehr, ob sein Schuß (im Vergleich mit anderen Schülern seines Alters) technisch „gut" oder vielleicht nur „ausreichend" ausgeführt worden ist, führt eine solche Transformation in ein vorher definiertes Bewertungssystem durch, und zwar in ein Bewertungssystem, das *nicht* durch den Bewertungsgegenstand selbst definiert ist. Die statistische Eigenschaft des Torverhältnisses ergibt sich aus den Spielregeln des Fußballspiels und ist z. B. für einen Schwimmwettkampf nicht möglich, während unser Zensurensystem nicht-gegenstandsspezifisch auf alle möglichen Handlungs-, Wissens- und Leistungsbereiche angewendet werden kann.

SCHRÖTER fährt fort: „Der gesamte Sport ist durchsetzt mit der Vergabe von Zensuren, deren Anforderungen man sich freiwillig unterwirft." Hier wird der Begriff „Zensuren" innerhalb eines Satzes gleich in zwei verschiedenen (und falschen) Bedeutungen benutzt: Mit der „Vergabe von Zensuren" kann einerseits nur das Ergebnis eines Leistungsvergleiches gemeint sein, nur daß diese Leistung (auch der Unterschied besser — schlechter zwischen zwei Wettkämpfen) eben nicht „vergeben" wird, sondern sich zwangsläufig aus der sportlichen Handlung selbst ergibt; andererseits scheint der Begriff „Zensuren" mit „Regeln, denen man sich freiwillig unterwirft", verwechselt zu werden.

Und weiter: „Ja, die gesamten Olympischen Spiele erhalten einen großen Teil ihrer Motivation durch die Möglichkeit, exakt und in Zahlen oder Rangfolgen ausdrückbar erreichte Leistungen benennen zu können" (S. 173 ff.).

Spätestens hier hätte dem Autor die von ihm selbst zurückgewiesene Verwechslung (s. o.) von Leistungsmessung („A ist 10.5 sec gelaufen. Mannschaft X hat gegen Y 2:0 gespielt") mit Notengebung (A ist „befriedigend" gelaufen, Mannschaft X hat trotz ihres Sieges nur „schwach ausreichend" gespielt) deutlich werden müssen.

Daß es sich hier nicht nur um eine zufällige Verwechslung der Begriffe handelt, zeigt sich auch noch an einer anderen Stelle. Bei der Frage, ob es möglich sei, im

Sportunterricht auf Zensuren zu verzichten, führt er die Aussage einer Lehrerin an: „Kinder wünschen eine Leistungskontrolle, die ihre Leistungen vergleichbar machen. Hierbei sind Zensuren für sie eindeutiger und leichter überschaubar als längere Beurteilungen" (S. 134).

Es muß schon eine arge Begriffsverwirrung herrschen, wenn man behauptet, Zensuren würden die Leistungen im Sport „eindeutiger" und „vergleichbarer" machen! Eine Leistungskontrolle findet im Sport (vielleicht ein wesentlicher Unterschied zu den anderen Fächern) in jeder Stunde, in jedem Bewegungsvollzug statt, im Wettlauf, im Korbwurf, in der gelungenen oder mißlungenen Kippe; Merkmal der sportlichen Vollzüge ist es eben, daß sie als Leistungen unmittelbar, objektiv und ohne Umweg über einen Dritten miteinander vergleichbar sind, daß sie unmittelbar, fachspezifisch und viel genauer differenzieren als unser Notensystem, das im Sport praktisch nur fünf Klassifikationen („sehr gut", „gut", „befriedigend", „ausreichend" und „mangelhaft") zuläßt. Nach einem 100-m-Lauf wissen die Schüler sehr genau, wer der schnellste Läufer der Klasse ist, die Stoppuhr liefert diese Information eindeutig und objektiv. Die Leistungen werden keineswegs leichter vergleichbar, wenn durch den Vorgang der Zensierung der Schüler, der 11,9 sec gelaufen ist, dieselbe Zensur erhält wie derjenige, der 12,0 sec benötigt hat.

Wenn es um Vergleich von Leistungen geht, ist die Zensur das denkbar ungeeignetste Mittel — und zudem völlig überflüssig.

Dieses Mißverständnis zeigt sich auch, wenn z. B. GÜNZEL (1975) schreibt: „Es ist zweifelsohne richtig, daß insbesondere die Sportnote dem einzelnen seine persönlichen Erfolge zu signalisieren (...) vermag" (S. 90). Der persönliche Erfolg oder Mißerfolg kann vom Schüler unmittelbar an der gelungenen oder mißlungenen Turnübung, am gelungenen oder mißlungenen Korbwurf, an der 100-m-Zeit abgelesen werden — die *Note* sagt dem Schüler nur, wie seine Leistung vom Lehrer nach einem Vergleich mit anderen eingeschätzt wird.

Gerade an diesen Begriffen „Erfolg und Mißerfolg" zeigt sich sehr deutlich die soziale Eingebundenheit der Zensur: Die Note kann ihm Mißerfolg signalisieren — Mißerfolg im Hinblick auf die Erwartungen des Lehrers, Mißerfolg im Vergleich zu anderen Schülern — obwohl der Schüler selbst vielleicht mit seiner Leistung höchst zufrieden ist und sie als Erfolg erlebt (und umgekehrt).

(Das Problem des Informationswertes der Zensur wird noch an anderer Stelle eingehender dargestellt.)

2.1 ALTERNATIV-ENTSCHEIDUNGEN

Als Beispiel für eine Art Zensur, mit der wir auch außerhalb der Schule konfrontiert sind, nennt SCHRÖTER den Führerschein und das Freischwimmerzeugnis.

M. E. handelt es sich hier um eine besondere Art von Prüfungen, um Alternativ-Entscheidungen, in die der Meß- und Bewertungsvorgang unmittelbar eingehen.

Alternativ-Entscheidungen der hier angesprochenen Art beruhen auf zwei Vorgängen:

a) *wird das betreffende Leistungskontinuum zwei Kategorien zugeordnet.* Bei der Prüfung zum Freischwimmerzeugnis: weniger als 15 Minuten — mehr als 15 Minuten schwimmen. Die Festlegung des kritischen Punktes (15 Minuten) ist letzten Endes willkürlich und beruht auf einer allgemeinen Vereinbarung; es wäre jederzeit möglich, die Grenze bei 14 oder 16 Minuten festzulegen.
Mit dieser Zuordnung zu zwei Kategorien entspricht die Alternativ-Entscheidung dem Zensieren mit dem einen Unterschied, daß die übliche Zensurenskala in sechs Kategorien differenziert, während hier nur zwei Kategorien (entweder-oder) zur Verfügung stehen.

b) *es wird mit dem kritischen Punkt zugleich eine rechtliche Konsequenz ausgedrückt*, die mit einer Zensur nicht notwendig verbunden ist, vielmehr in einem zusätzlichen Entscheidungsakt ausdrücklich hinzugefügt werden muß.

So geben z. B. alle Prüfungsordnungen zum einen die zu verwendende Notenskala an, und in einem weiteren Paragraphen wird festgelegt, welcher Notendurchschnitt erreicht werden muß, damit die Prüfung bestanden ist, d. h. es wird ein kritischer Punkt angegeben, der eine Alternativ-Entscheidung definiert.
Noch ein weiterer — allerdings nicht notwendiger — Unterschied besteht zwischen der Zensierung nach einer Notenskala und nach dem Bestanden-Nicht-Bestanden-Prinzip:
Mit der (+ —)-Bewertung ist — wie oben bereits angedeutet — meist eine rechtliche Konsequenz verbunden: Der Fahrschüler bekommt den Führerschein oder eben nicht; der Schüler darf sich Freischwimmer nennen oder nicht. Diese Alternativ-Entscheidung hat zur Folge, daß der kritische Punkt meist sehr genau bestimmt, die Prüfung *operationalisiert* wird: Freischwimmer ist, wer 15 Minuten ohne Unterbrechung schwimmt, egal ob er 16 oder 25 Minuten durchhält, er muß nur den kritischen Punkt erreichen, und im Hinblick auf „pädagogische Zensierung": es interessiert nicht, warum der Schüler vielleicht nur 10 Minuten durchhält — das ist zwar bedauerlich, aber er ist halt noch kein Freischwimmer. D. h. in der Tendenz ist eine Alternativ-Entscheidung (eben durch die leichtere Operationalisierbarkeit) *unpersönlicher im Hinblick auf den Schüler* (er wird nicht nach „sehr gut", „gut" ... abgestuft qualifiziert) und *unpersönlicher im Hinblick auf den Lehrer* (durch die Operationalisierung ist meist allein durch den Meßvorgang die Entweder-Oder-Entscheidung schon gefallen: die Uhr zeigt an, ob der Schüler Freischwimmer ist oder nicht).
Hinzu kommt, daß Alternativ-Prüfungen meist *einmalige Ereignisse* darstellen: wer einmal das Freischwimmerzeugnis erworben hat, braucht sich nicht noch einmal darum zu bemühen, d. h. er steht nicht in einem andauernden Bewertungsprozeß, der eine kontinuierliche Bewährung von ihm verlangt, wie es normalerweise mit unserem Zensurensystem verbunden ist.

Es läßt sich zusammenfassend sagen:

Alternativ-Entscheidungen sind eine Sonderform der Zensuren: Die Schüler werden aufgrund ihrer Leistungen nach zwei Kategorien qualifiziert; der kritische Punkt ist willkürlich gesetzt; der Qualifizierung geht ein Meßvorgang voraus.

Die Alternativ-Entscheidungen unterscheiden sich (praktisch, nicht notwendigerweise) dadurch von den üblichen Zensuren, daß mit der Entscheidung meist unmittelbar eine rechtliche Konsequenz verbunden ist; daß der kritische Punkt meist genau operationalisiert ist; daß mit dem Meßvorgang automatisch auch die Qualifikation entschieden, die Entweder-Oder-Entscheidung gefällt wird.

Die Entweder-Oder-Entscheidung als primitivste Form der Zensurenskala mit nur zwei Kategorien ist überall dort einzusetzen, wo die Qualifizierung der Schüler nicht aufgrund eines Vergleichs zwischen den Schülern erfolgt sondern aus *sachlichen Notwendigkeiten* her vorgenommen wird.

Um das zu verdeutlichen: Ein „befriedigend" in einem Schulfach wird normalerweise für eine Leistung vergeben, die im Durchschnitt von den Schülern der betreffenden Altersstufe erbracht wird. D. h. der Lehrer muß wissen, was normalerweise von einem Schüler verlangt werden kann, wenn er dessen Leistungen beurteilen soll. So legen die KMK-Vereinbarungen z. B. für eine „3" im Schwimmen eine bestimmte Leistung über 50 m Kraul für Jungen fest. D. h. einer Zensierung geht (bewußt oder unbewußt) immer erst eine Leistungsmessung in einer größeren Population voraus, aufgrund derer die Einzelleistung des Schülers A eingeordet und bewertet wird.

Bei Alternativ-Bewertungen wird normalerweise anders vorgegangen: man versucht — möglichst sachlich — festzustellen, welche Kenntnisse ein Fahrschüler haben muß, damit er ein Fahrzeug führen darf — unabhängig davon, wie viele Personen (und unter welchem Aufwand) diese Kenntnisse erwerben. Oder: es wird als erstrebenswertes Ziel festgelegt, daß jeder Bürger wenigstens 15 Minuten lang schwimmen können sollte — unabhängig davon, wie viele Freischwimmer es bereits gibt, im Gegenteil: es ist das erklärte Ziel, daß möglichst alle die kritische Grenze von 15 Minuten erreichen.

(Man könnte natürlich sagen, es sei ebenfalls das erklärte Ziel der Sportlehrer, daß alle Schüler die Leistungen für eine „1" erreichen. Das ist aber im Idealfall nur in kleinen Gruppen, z. B. innerhalb einer Klasse möglich, da die Zensur auf dem interindividuellen Vergleich immer vorhandener Leistungsunterschiede beruht.)

Es gibt noch eine Sonderform der Alternativ-Zensur, die dadurch entsteht, daß die Menge derer, die den kritischen Punkt überschreiten dürfen, überschreiten sollen, aus sachlichen Zwängen begrenzt ist. Damit ist folgendes gemeint: Im NC wird festgelegt, wie viele Studienplätze in einem Fach vorhanden sind, und so viele Studenten werden zugelassen. Der kritische Punkt, die notwendige Durchschnittsnote, ergibt sich aus dem NC und dem Notendurchschnitt aller Bewerber.

Der Lehrer, der eine Schulmannschaft im Fußball aufstellen soll, kann nach eingehender Prüfung der in Frage kommenden Schüler maximal elf Schüler nominieren, unabhängig davon, ob er an einer Schule für Körperbehinderte oder an einem Sportgymnasium unterrichtet.

D. h. in solchen Fällen unterscheidet sich die Alternativ-Zensur von der normalen Zensur wesentlich dadurch, daß der kritische Punkt notwendig *nicht* willkürlich festgelegt wird, sondern sich aus anderen, vorhergehenden Entscheidungen ergibt (z. B. aus der Entscheidung, daß eine Fußballmanschaft aus elf Spielern besteht). Genaugenommen genügt es in solchen Fällen, die Schüler/Bewerber in eine Rangreihe zu bringen, was allerdings wieder — wie bei der üblichen Zensur — einen interindividuellen Vergleich verlangt, wenn auch nur nach Besser/Schlechter-Relationen. Das führt natürlich, wie die Entwicklung an unseren Oberschulen unter dem Einfluß des NC in den letzten Jahren gezeigt hat, zu einem Kampf aller gegen alle: in einem Rangsystem ist eben die Verbesserung eines individuellen Rangplatzes grundsätzlich nur auf Kosten eines Konkurrenten möglich.

Leistungsmessungen, die mit einer Alternativ-Entscheidung verbunden sind, stellen einen Sonderfall einer kriterium-orientierten Messung dar. Der Unterschied besteht darin, daß in der kriterium-orientierten Messung der Abstand des Schülers vom vorher definierten Lernziel festgestellt wird, während hier nicht der Abstand vom Lernziel interessiert, sondern nur, ob die kritische Grenze (z. B. die 15 Minuten beim Freischwimmer) unterschritten oder erreicht wird. Wie weit der Schüler ggf. vom Ziel entfernt ist, interessiert erst in 2. Linie, wenn er die 15 Minuten nicht erreicht, ist er nicht Freischwimmer, gleichgültig, ob er das Ziel um fünf Sekunden oder um zehn Minuten verfehlt hat.

Im Unterschied zur kriterium-orientierten Leistungsmessung wird die mit dem Erreichen des Ziels verbundene Qualifikation genau wie eine Zensur durch einen Dritten zugeteilt. Um Freischwimmer zu sein, genügt es nicht, 15 Minuten zu schwimmen. Diese Leistung muß auch noch von einer offiziellen Instanz festgestellt und die Qualifikation „Freischwimmer" ausgesprochen werden. (Ähnlich wird der Führerschein zu- oder aberkannt, ein Examen setzt eine offizielle Prüfungsinstanz voraus etc.)

2.2 DAS SKALENNIVEAU VERSCHIEDENER MESSMETHODEN

In Kapitel 1 wurde der Unterschied zwischen Messen und Bewerten aufgezeigt. Nun muß das Messen selbst etwas differenzierter dargestellt werden. Oben wurde gesagt: Messen besteht in der Zuordnung von Zahlenwerten mit dem Anspruch, daß der zugeordnete Zahlenwert das gemessene Objekt in einem vorher bestimmten Zahlenraum eindeutig abbildet, ihm einen bestimmten Platz zuweist.

Nun gibt es mehrere verschiedene Zahlenräume, in denen eine solche Platzzuweisung erfolgen kann. Welcher Raum benutzt wird, hängt u. a. von den Informationen ab, die wir über das zu messende Objekt (oder Verhalten) gewinnen können. Entsprechend hat jeder Zahlenraum bestimmte mathematische Eigenschaften (die nur bestimmte mathematische Operationen zulassen) und liefert auch nur bestimmte Informationen über das gemessene Merkmal.

Da sich eine Zeugnisnote normalerweise aus vielen einzelnen Meßdaten aus vielen verschiedenen Leistungs- und Verhaltensbereichen ergibt, ist es notwendig, sich ganz kurz mit den wichtigsten Meßverfahren („Skalen") zu beschäftigen.

2.2.1 Nominalskala

Die einfachste und auch häufigste Form des Messens ist das Zählen. (Der Begriff „messen" muß hier etwas weiter als im alltäglichen Sprachgebrauch im Sinne von „quantifizieren" verstanden werden.)

Das Zählen setzt voraus, daß wir ein bestimmtes Merkmal definieren, das in verschiedenen Ausprägungen auftreten kann. Das Merkmal, das wir messen wollen, kann z. B. „Geschlecht" sein, die verschiedenen Ausprägungen wären „männlich" und „weiblich". Durch den Vorgang des Zählens können wir z. B. feststellen, daß in einer Klasse 12 Jungen und 18 Mädchen sind. Mit diesen beiden Meßzahlen können wir dem Meßgegenstand „Klasse" hinsichtlich des Merkmals „Geschlechtsverteilung" einen ganz bestimmten Platz zuordnen.

Die Nominalskala liefert als Information Häufigkeiten, also wie oft eine bestimmte Merkmalsausprägung auftritt.

Das Beispiel „Jungen und Mädchen" zeigt auf triviale Weise, daß in diesem Zahlenraum die Meßwerte keinen weiteren mathematischen Operationen unterworfen werden können. Es wäre unsinnig, etwa einen Mittelwert bilden zu wollen, da die beiden Merkmalsausprägungen „männlich — weiblich" einander ausschließen. Wir können die Meßwerte auch nicht addieren, es sei denn, wir bilden damit eine neue Merkmalskategorie „Schüler": die Klasse A hat 30 Schüler, Klasse B hat X Schüler etc.

Für eine Nominalskalierung brauchen wir also mehrere Ereignisse (Häufigkeiten) in diskreten Merkmalsausprägungen, die aber qualitativ gleich sind. Über die einzelnen Ereignisse gibt es keine weitere Information.

Im Hinblick auf Messen und Bewerten im Sportunterricht spielt die Nominalskalierung kaum eine Rolle, da sie eben über das einzelne, individuelle Ereignis, das dann zur Zensur führt, nichts aussagt. Wir benutzen die Nominalskalierung, wenn wir z. B. am Ende einer Unterrichtsreihe prüfen, wie viele Schüler das Lerziel erreicht haben und wie viele nicht: 20 Schüler können die Kippe, 12 können die Kippe nicht.

2.2.2 Rangskala

Die nächsthöhere Skala ist die *Rangskala*. Die Rangskalierung kann immer dort angewendet werden, wo das zu messende Merkmal in verschieden starker Ausprägung auftritt und der Grad der Ausprägung nur nach größer, kleiner oder gleich festgestellt wird.

„Die Sachverhalte, die einer Ordinalskala zugänglich sind, *gleichen* einander im Hinblick auf das betrachtete Merkmal, sie *unterscheiden* sich jedoch voneinander bezüglich der Merkmals*ausprägung* (Größe, Stärke, Intensität). Es bleibt undefiniert, wie stark die Differenzen zwischen den verschiedenen Objekten sind" (Claus / Ebner, 22).

Wenn wir die Schüler fragen, was sie lieber spielen, Fußball oder Handball, dann nehmen sie eine Rangskalierung vor, wenn sie sagen „Fußball lieber als Handball". Fußball erhält den Rangplatz 1 (den Meßwert 1), Handball den Rangplatz

2. Mit diesen Meßwerten ist nichts darüber ausgesagt, um wieviel die Schüler lieber Fußball spielen.

Wenn drei Schüler um die Wette laufen (ohne daß der Lehrer die Zeiten stoppt), ergibt sich, daß A schneller war als B, und B schneller als C. A erhält den Rangplatz 1, B Rangplatz 2, C Rangplatz 3. Mit diesen Meßwerten ist nichts darüber ausgesagt, um wieviel jeweils A schneller war als B, und B schneller war als C. A könnte weit vor B ankommen, während B nur ganz knapp C schlägt; es könnte aber auch A nur ganz knapp vor B sein, während C weit abgeschlagen ist.

Die Eigenschaft der Rangskala, daß der *Abstand* zwischen verschiedenen Rängen *nicht definiert* ist, schließt bestimmte Rechenvorgänge mit Rang-Meßdaten aus. So ist es z. B. nicht möglich, aus verschiedenen Rang-Meßdaten desselben Individuums einen Mittelwert zu bilden — ein Fehler, der in der Praxis sehr häufig auftritt, und der deshalb hier an einem Beispiel aus dem Sport verdeutlicht werden soll.

Drei Schüler (A, B, C) vergleichen sich im 100-m-Lauf. Sie machen drei Durchgänge. Da sie keine Stoppuhr besitzen, stellen sie in jedem Durchgang nur fest, wer 1., 2. oder 3. war. Sie erhalten folgende Tabelle.

	1. Lauf	2. Lauf	3. Lauf
Schüler A	1.	2.	2.
Schüler B	2.	1.	1.
Schüler C	3.	3.	3.

Wenn sie feststellen wollen, wer denn nun eigentlich der beste Läufer sei, so ist das nur möglich, indem sie auszählen, wer am häufigsten Rangplatz 1, wer dann am häufigsten Rangplatz 2 etc. hat. B wäre 1., A wäre 2. und C 3.

Würden wir, um „exakter" zu sein, einen Mittelwert für jeden Schüler errechnen, so erhielten wir für

Schüler A 1.66
Schüler B 1.33
Schüler C 3.00

Die unterschiedliche Größe dieser „Mittelwerte" spiegelt zwar das Resultat wider, das wir aufgrund der Häufigkeiten errechnet haben (B = 1., A = 2., C = 3.), die numerische Größe dieser Werte ist aber in zweifacher Hinsicht unsinnig:

1. gibt es keinen „Eins-Komma-sechsten Platz — Plätze sind immer ganzzahlig,
2. täuscht die numerische Größe eine Information vor, die dieser „Mittelwert" gar nicht geben kann:

Der Unterschied zwischen A und B beträgt 0.33, zwischen A und C 1.34. Der Schluß, daß A etwa viermal so weit von C entfernt ist wie B von A, daß also A viel näher bei B liegt als bei C, ist falsch bzw. kann aus diesen Daten nicht gezogen werden. Verdeutlichen wir uns das wieder an unseren drei Schülern, die um die Wette laufen:

Nehmen wir an, der Lehrer habe die drei Durchgänge doch gestoppt und folgende Werte erhalten:

	1. Lauf	2. Lauf	3. Lauf	Mittelwert
Schüler A	10,5	12,1	12,1	11,6
Schüler B	12,0	12,0	12,0	12,0
Schüler C	12,2	12,2	12,2	12,2

Nun zeigt sich plötzlich, daß sich aufgrund des Mittelwertes der gestoppten Zeiten die Reihenfolge A > B > C ergibt, daß A auch absolut die beste Zeit gelaufen ist, und daß in der durchschnittlichen Leistung B näher bei C als bei A liegt.

Wir sehen also, daß eine Rangskala nur sehr ungenaue Informationen gibt, da die Abstände zwischen den einzelnen Meßwerten nicht definiert ist, und daß die Bildung eines Mittelwertes zu völlig unsinnigen Schlüssen führt.

Die Rangskalierung spielt überall dort eine Rolle, wo für das zu messende Merkmal keine sinnvollen Meßeinheiten definiert werden können.

Wenn wir z. B. den Leistungswillen eines Schülers feststellen wollen, dann können wir wohl sagen: „A strengt sich mehr an als B". Es gibt aber kaum eine Möglichkeit zu sagen: „A strengt sich um X Meßeinheiten mehr an als B." (In der Psychologie wird viel Mühe darauf verwandt, auch für solche Merkmale Meßsysteme zu entwickeln, die Aussagen dieser Art zulassen. Die Darstellung dieser Methoden würde aber unseren Rahmen hier sprengen.)

Im sportmotorischen Bereich findet eine Rangskalierung z. B. in Wettkämpfen nach dem K.o.-System statt oder in Turnieren „jeder gegen jeden". Die Endplacierung in einem Boxturnier sagt nichts darüber aus, um wieviel der Sieger besser war als der Zweitplacierte.

Auf eine Eigenart der Rangskalierung muß noch deutlich hingewiesen werden: Wenn ein Individuum A seinen Rangplatz *verbessert*, muß sich automatisch ein anderes Individuum B *verschlechtern* (wenn wir von dem Sonderfall der gleichen Platzziffer absehen); obwohl also die Leistung von B konstant geblieben ist, bekommt B eine andere Meßzahl zugeordnet, weil A seine Leistung geändert hat. Es ist deshalb mit Vorsicht zu genießen, wenn Günzel (1975) schreibt: „Ermunterung und Ansporn haben leistungsschwache Schüler deshalb besonders nötig, weil für sie Auszeichnungen im Fach Sport eminent wichtig sind im Hinblick auf ihr Bemühen, einen aktuellen Status innerhalb ihrer Klasse oder Gruppe zu erringen oder ihren schlechten Rangplatz zu verbessern, ein Motiv, das übrigens auch von Schülern recht häufig genannt wird" (S. 91). In einem *Rang*system ist immer einer der letzte.

Das Problem der Rangskalierung wurde deshalb hier so ausführlich dargestellt, weil angenommen werden kann, daß unsere Notenskala weitgehend einer Rangskala entspricht. Die Probleme, die sich daraus ergeben, werden weiter unten dargestellt.

2.2.3 Intervallskala

Die nächsthöhere Skala bezeichnet man als *Intervallskala.*

In dieser Skala können wir nicht nur feststellen, daß A größer (besser, schneller, schöner . . .) ist als B, sondern auch, um *wieviel* A sich von B unterscheidet; das läßt auch die Aussage zu, daß sich A um ebenso viel von B unterscheidet wie C von D, d. h. in dieser Skala sind die „Abstände zwischen den benachbarten (aufeinanderfolgenden) Skalenwerten konstant" (Claus / Ebner 1975, 23).

Klassisches Beispiel für eine Intervallskala ist das Thermometer nach Celsius: Wenn drei Gegenstände die Temperaturen A = 30 Grad Celsius, B = 20 Grad Celsius und C = 10 Grad Celsius haben, so ist die Aussage zulässig und richtig, daß der Unterschied zwischen A und B ebenso groß ist wie zwischen B und C. Wir können aber nicht sagen, 20 Grad Celsius sei doppelt so warm wie 10 Grad Celsius, diese Aussage ist unzulässig, da der Nullpunkt der Celsius-Skala willkürlich gesetzt ist, während der absolute Nullpunkt bei —273 Grad Celsius liegt. D. h. die Aussage —253 Grad ist doppelt so warm wie —263 Grad ist richtig, weil hier der absolute Nullpunkt berücksichtigt wird.

Eine Entsprechung im Sport finden wir z. B. in der Zehn-Kampf-Wertung: Wer im Weitsprung 400 Punkte erhält, springt keineswegs doppelt so weit wie jemand, der nur 200 Punkte erzielt.

Wer für seine Turnübung 10 Punkte erhält, turnt nicht doppelt so gut wie jemand, der nur 5 Punkte bekommt.

In beiden Fällen ist der Nullpunkt willkürlich gesetzt, die Messung erfolgt (annähernd) intervallskaliert.

2.2.4 Verhältnisskala

Das ändert sich in der höchsten Skalenform, der *Verhältnisskala.* Diese besitzt einen natürlichen Nullpunkt.

Am deutlichsten wird die Verhältnisskala am Beispiel des Metermaßes. Die Maße 10 cm, 20 cm, 30 cm lassen nicht nur die Aussage zu, daß die Differenz von 10 und 20 cm ebenso groß ist wie die von 20 und 30 cm, sondern es ist auch richtig und sinnvoll zu sagen, daß 20 cm doppelt so viel ist wie 10 cm.

Verhältnisskalierte Meßwerte lassen alle mathematisch-statistischen Rechenoperationen zu.

Die Verhältnisskala spielt bei der Messung von Sportleistungen eine wesentliche Rolle, und zwar werden alle C-G-S-Sportarten (Lauf, Sprung, Wurf, Gewichtheben etc.) in diesem System abgebildet.

Im Gegensatz zu Ballreich (In: Rieder 1972) bin ich der Meinung, daß auch Spielleistungen, die durch Punkte gezählt werden (z. B. Fußball, Tennis etc.) verhältnisskaliert sind, da hierbei der Ausprägungsgrad desselben Merkmals variiert, und die Angabe eines natürlichen Nullpunktes sinnvoll ist. Diese Häufigkeiten (Tore, Punkte) lassen im Gegensatz zur Nominalskala alle Rechenoperationen zu.

Vergleichen wir also nochmals die drei wichtigsten Skalentypen:

die *Rang-* (oder *Ordinal-*) *Skala* sagt, *daß* A besser ist als B;
die *Intervallskala* sagt, um *wieviel* A besser ist als B;
die *Verhältnisskala* gibt an, in *welchem Verhältnis* A besser ist als B.

Oder: Ein Zehn-Kämpfer kann im Bereich der Verhältnisskala sagen: „Ich bin doppelt so weit gesprungen wie B.";
im Bereich der Intervallskala: „Ich habe im Weitsprung 200 Punkte mehr als B." und im Bereich der Rangskala: „Im Weitsprung lag ich vor B."
Wir sehen an diesem Beispiel: die höherwertige Skala kann — unter Informationsverlust — in die nächstniedrigere überführt werden, während der umgekehrte Weg nicht möglich ist: Wenn wir nur wissen, daß A weiter gesprungen ist als B, können wir daraus nicht ableiten, *um wieviel* oder gar *in welchem Verhältnis* A besser war als B.
Ferner können wir festhalten: Veränderungen in der Intervall- und der Verhältnisskala sind absolut (individuell) möglich (A springt 1 m weiter oder doppelt so weit wie vorher), während Veränderungen in einer Rangskala immer relativ sind (wenn B sich um einen Rangplatz verbessert, muß sich A um einen Platz verschlechtern; s. o.).
Die Bedeutung dieser Überlegungen zum Skalenniveau verschiedener Meßdaten für unser Thema wird sofort deutlich, wenn wir uns vergegenwärtigen, daß eine Zeugniszensur in fast allen Fällen von Meßwerten abgeleitet werden, die verschiedenes Skalenniveau besitzen: Leichtathletikleistungen sind verhältnisskaliert, Turnleistungen intervallskaliert, Beteiligung am Unterricht rangskaliert, während die Feststellung „er ist Freischwimmer" am ehesten der Nominalskala entspricht. Deshalb haben — genaugenommen — Zeugniszensuren allenfalls den Informationswert von Rangskalen (vgl. Lienert 1962), und das würde z. B. die Bildung von Mittelwerten (wie es bei den Berechnungen zum NC geschieht) verbieten. (Auf dieses Problem werden wir noch unten — Kapitel 8 — zu sprechen kommen.)

Messen verschiedener Variablen

Im Sport gibt es einige Sonderfälle, in denen die Leistung mit zwei verschiedenen Meßsystemen erfaßt wird:
Beim Skispringen z. B. wird Haltung u n d Weite zu einem Gesamtwert zusammengefaßt: die Haltungsnote ist intervallskaliert (es gibt keinen natürlichen Nullpunkt, denn es gibt nicht „gar keine Haltung"), während die Weite Verhältnisskalenniveau besitzt (das Metermaß hat einen natürlichen Nullpunkt). Das Ergebnis ist deshalb auch nur auf der Intervallskala abgebildet.
Bei den Wertungssportarten werden Schwierigkeit und Ausführung zusammengefaßt, wobei die Schwierigkeit der Übung vorher bereits intervallskaliert festliegt: ein Sprung, der den Schwierigkeitsgrad 2.0 besitzt, ist nicht doppelt so schwer wie ein Sprung mit der Schwierigkeit 1.0; denn es gibt keinen natürlichen Nullpunkt, keinen Sprung, der überhaupt keine Schwierigkeit hat. Die Bewertung der Ausführung durch die Punktrichter ist ebenfalls — wie beim Skispringen — intervall-

skaliert. Probleme treten dadurch auf, daß hier zwei qualitativ verschiedene Meßobjekte (Schwierigkeit und Ausführung) — auch wenn sie auf demselben Niveau skaliert sind — in einem Wert zusammengefaßt sind (Aus der Grundschule wissen wir, daß nur gleichartige Dinge addiert werden dürfen).
Die Probleme, die sich daraus für die Zensierung ergeben, werden in Kapitel 6 eingehender dargestellt.

3. Die Kriterien für das Messen von Leistungen

Oben wurde bereits festgestellt: Jeder Bewertung muß ein Meßvorgang vorausgehen. Nun müssen wir noch erweitern: bevor wir messen, müssen wir genau definieren, *was* wir messen wollen, und *wie* wir es messen wollen.

Hinsichtlich dessen, *was* gemessen werden soll, sind sich die Sportpädagogen weitgehend uneinig. Während kein Zweifel darüber besteht, daß die objektive (sportmotorische) Leistung gemessen werden und in die Zensur eingehen soll, ist man sich auch noch darüber weitgehend einig, daß *nicht nur* diese objektive Leistung berücksichtigt werden soll. Dagegen gehen die Meinungen darüber, welche Verhaltensweisen noch (z. B. Leistungswille, Mitarbeit, soziales Verhalten, Kreativität, Veranlagung etc.), und in welchem Gewichtungsverhältnis diese zur motorischen Leistung gemessen und berücksichtigt werden sollen, weit auseinander (vgl. Tabelle 6, Seite 78).

Daß diese nicht-motorischen Merkmale in der Zensur zu berücksichtigen sind, ist auch in den Rahmenrichtlinien zum Sportunterricht der meisten Bundesländer amtlich festgehalten, ohne daß dabei angegeben ist, *wie* sie gemessen und *mit welchem Gewicht* sie berücksichtigt werden sollen.

Im sportmotorischen Bereich ist die Definition des Meßobjekts relativ einfach und meist durch die Wettkampfregeln fesgelegt: im 100-m-Lauf ist es die Zeit, im Kugelstoßen die Weite, im Gewichtheben das zur Hochstrecke gebrachte Gewicht, im Turnen die Kombination von Schwierigkeit und Ausführung, im Schießen die getroffenen Ringe etc.

Diese Definitionen sind nicht so selbstverständlich wie sie uns auf den ersten Blick erscheinen. Der Gedanke, beim 100-m-Lauf nicht die Zeit sondern die Haltung zu messen, kommt uns absurd vor — aber das liegt nicht „in der Natur der Sache" — sondern ist das Ergebnis einer Vereinbarung. So wird z. B. beim Skispringen nicht nur die Weite (wie beim Weitsprung) sondern auch die Haltung gewertet (= gemessen).

Vor einigen Jahren gab es im Rudersport vorübergehend für Frauen die Disziplin „Stilrudern"; hierbei wurde nicht die Zeit, die ein Boot für die vorgeschriebene Strecke benötigte, gemessen, sondern der Rhythmus und die Harmonie der Mannschaft. (Diese Wettkampfart wurde bald wieder fallengelassen, u. a. vielleicht aufgrund der Tendenz, auch die Leistungen im Sport auf einem möglichst hohen Skalenniveau zu messen.)

Im alpinen Skilauf wird normalerweise die benötigte Zeit als zu messende Leistung bestimmt (also auf dem Niveau der Verhältnisskala), während seit einiger Zeit auch Wettkämpfe im „Parallelslalom" ausgetragen werden, in dem die Leistungen (im K.-o.-System) rangskaliert gemessen werden: Nicht mehr die benötigte Zeit ist wesentlich, sondern die Frage, wer als 1 und wer als 2 angekommen ist.

Im Gegensatz zu den sportmotorischen Leistungen sind Verhaltensweisen wie Mitarbeit, Leistungswille, soziales Verhalten etc. kaum operationalisierbar, ja nicht einmal genau definierbar. Was verstehen wir unter „sozialem Verhalten"? Von zehn Lehrern werden wir vielleicht zehn verschiedene Antworten auf diese Frage erhalten. Noch schwieriger wird es sein, für ein so komplexes Merkmal eine angemessene Meßmethode zu entwickeln: wie messen wir, in welchem Ausmaß sich jemand sozial verhält? (Ein großer Teil der Psychologen befaßt sich eben mit derartigen Problemen.)

„Und wenn man schon den Charakter, das sportliche oder unsportliche Verhalten des Schülers in der Zensur deutlich werden lassen möchte, so ist nach dem Kriteriensystem zu fragen, mit dessen Hilfe „Charakter" bewertet werden kann. Es ist bis heute noch nicht vorhanden" (SCHRÖTER 1977, 126).

Mit dieser Frage nach der Definition des Meßobjekts und des Meßvorgangs befinden wir uns in dem Gebiet, das sich in der *Objektivität*, der *Reliabilität* und der *Validität* befaßt.

Es ist notwendig, die Leistungsmessung im Sport, vor allem wenn sie im Hinblick auf Zensuren erfolgt, daraufhin zu untersuchen, wie weit sie diese Kriterien erfüllt. (Wettkampf und Leistungsmessung entsprechen in wesentlichen Merkmalen dem psychologischen Test, den LIENERT (1967, 7) definiert als „wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung.")

Beschränken wir uns zuerst auf das Messen sport*motorischer* Leistungen.

3.1 OBJEKTIVITÄT

„Unter Objektivität eines Testes verstehen wir den Grad, in dem die Ergebnisse eines Testes unabhängig vom Untersucher sind" (LIENERT 1967, 13). Das bedeutet für die Leistungsmessung im Sport: Verschiedene Beurteiler/Kampfrichter/Lehrer müssen bei einer Leistungsmessung/Wettkampf/Prüfung zu möglichst demselben Ergebnis kommen.

Die Bedeutung dieses Kriteriums „Objektivität" ist klar: bei niedriger Objektivität hängt die Zensur, die der Schüler bekommt, weniger von seiner Leistung ab als vielmehr davon, welcher Lehrer seine Turnübung (seinen Deutschaufsatz, seine Vokabelkenntnisse) bewertet (vgl. INGENKAMP 1972).

Die Objektivität der Leistungsmessung ist in den einzelnen Sportarten verschieden hoch; die Meßverfahren selbst sind — wenn sie sich nicht trivialerweise von selbst ergeben — in den Spiel- und Wettkampfregeln festgelegt.

Für einige Sportarten ist (vor allem bei großen Wettkämpfen durch die Einführung elektronischer Meßmethoden) praktisch vollständige Objektivität gegeben: in allen C-G-S-Sportarten ist die Leistung völlig unabhängig vom Lehrer (oder Kampfrichter), einzige Voraussetzung ist, daß er z. B. das Bandmaß richtig abliest. Interindividuelle Unterschiede in der Reaktionszeit, z. B. beim Bedienen einer Stoppuhr, können durch elektronische Zeitnahme ausgeschaltet werden. Prinzipiell ist hier — abhängig vom technischen Einsatz — eine beliebig große Genauigkeit und eine vollständige Objektivität möglich.

Wir dürfen dabei natürlich nicht außer acht lassen, daß in der konkreten Schulsituation diese Objektivität oft wesentlich niedriger ist. Wenn z. B. bei Schulsportfesten Nicht-Sportlehrer als Kampfrichter fungieren, kommt es häufig zu völlig falschen Meßwerten, weil das Maßband an der falschen Stelle angelegt oder die Stoppuhr im falschen Zeitpunkt gedrückt wird. Das sind jedoch Fehler, die nicht grundsätzlich im Meßinstrument begründet sind und durch eine genaue Einweisung des Benutzers vermieden werden können.

Beim Kegeln läßt sich eindeutig feststellen, ob ein Kegel gefallen ist oder nicht; beim Schießen gibt es kaum eine Frage, ob (oder wie oft) der Schütze getroffen hat etc.

Wesentlich geringer ist die Objektivität der Leistungsmessung in denjenigen Sportarten, deren Leistungen keinen *Produktcharakter*, der in *Quantitäten* ausgedrückt werden kann, besitzt (100 m in 12 sec), sondern die *Prozeßcharakter* haben und nach *qualitativen* Merkmalen bewertet werden (für die Exaktheit der Ausführung erhält der Turner 9.30 Punkte). Diese Qualitäten werden vom Lehrer/Kampfrichter aufgrund eines subjektiven Eindrucks gewertet. (In der Psychologie würde man diese Art der Leistungsmessung als „rating" bezeichnen.)

Diese grundsätzliche Subjektivität kann niemals völlig ausgeschlossen werden. Wohl aber kann durch verschiedene Maßnahmen ihr Einfluß auf den Meßvorgang vermindert (die Objektivität erhöht) werden:

a) Es werden bei größeren Wettkämpfen mehrere Kampfrichter eingesetzt und deren Punktwertungen gemittelt, wobei (z. B. beim Wasserspringen) möglicherweise Extremwerte nicht berücksichtigt werden. Man geht hier von der Annahme aus, daß vier Augen mehr sehen als zwei, daß sich verschiedene subjektive Fehler gegenseitig aufheben, d. h. daß der Mittelwert vieler verschiedener Wertungen der „Wahrheit" am nächsten kommt.
b) Es werden genaue Wertungsvorschriften erarbeitet, mit denen versucht wird, so wenig operationalisierbare Leistungsmerkmale wie Eleganz, Originalität, Harmonie, Übereinstimmung mit der Musik etc. möglichst detailliert zu erfassen und im Hinblick auf die Gesamtwertung zu gewichten. (Ein gutes Beispiel für eine solche differenzierte Wertungsvorschrift ist der „Code de pointage" für die Wettkampfgymnastik.)
c) Vor wichtigen Wettkämpfen werden die Kampfrichter einem intensiven Training unterzogen im Hinblick darauf, was als Fehler gilt, wofür es Punktabzug gibt, und wie hoch dieser sein soll, um die Übereinstimmung zwischen den Kampfrichtern zu erhöhen.

   Selbstverständlich kann mit diesen Mitteln nur erreicht werden, daß (bei qualifizierten und um Objektivität auch wirklich bemühten Kampfrichtern!) der Einfluß subjektiver Faktoren herabgesetzt (die Objektivität erhöht) wird.

   Die Objektivität des „Ratings" hängt unlösbar vom guten Willen des jeweiligen Kampfrichters ab; „Sympathiewertungen" oder ganz offene Bevorzu-

gung aufgrund der Nationalität des Wettkämpfers und des Kampfrichters (wie sie immer wieder in den „Bewertungsskandalen" beim Kunstturnen und besonders beim Eiskunstlauf auftreten) sind nie völlig auszuschließen. Eine Objektivität wie in den C-G-S-Sportarten ist keinesfalls erreichbar.
Wenn das Problem des subjektiven Einflusses sogar durch diese, z. T. sehr aufwendigen Mittel im Leistungssport nicht oder nur unzureichend gelöst werden kann, so dürfen wir für den Schulalltag ruhig annehmen, daß eine Turnübung, die von Lehrer A mit fünf Punkten bewertet (= gemessen) wird, von Lehrer B möglicherweise mit sechs oder nur mit vier Punkten bedacht wird.

Die Objektivität der Leistungsmessung in den „Kunst"-Sportarten wird auch dadurch beeinträchtigt, als in die Messung zwei Faktoren mit unterschiedlichem Meßniveau eingehen: es wird nicht nur die Ausführung, sondern ebenfalls die Schwierigkeit der Übung bewertet. Dieses Problem wird dadurch weitgehend ausgeräumt, daß jedem Übungsteil ein bestimmter Schwierigkeitsgrad zugeordnet wird, der dann (in verschiedenen Sportarten in unterschiedlicher Weise) mit der Ausführung verrechnet wird. Bei dieser Art der Verrechnung wird allerdings vorausgesetzt, daß der Schwierigkeitsgrad einer Übung unabhängig ist von seiner Ausführung, was ja nicht ganz stimmt: ein gestreckter Salto mit krummen Knien — also mit schlechter Ausführung — ist objektiv leichter als eine gute Ausführung in völliger Streckung. Umgekehrt ist ein schlecht gehockter Salto schwerer, da er langsamer dreht, als ein eng gehockter, „gut" ausgeführter Salto. Das Problem, das hier auftaucht, wird dadurch praktisch gelöst, daß man durch Übereinkunft den Schwierigkeitsgrad unabhängig von der Ausführung konstant hält.

Das Problem der Objektivität bei der Leistungsmessung in Spielen muß unter zwei verschiedenen Aspekten betrachtet werden, je nachdem, was als Leistung gemessen werden soll.
Zum einen ist in den Wettkampfregeln die Leistung einer Mannschaft definiert durch die Anzahl der Tore (Körbe, Punkte etc.), die sie erzielt.
Bei der Feststellung *dieser* Leistung kommt es häufiger zu Zweifelsfällen, ob ein Tennisball die Auslinie, der Fußball schon die Torlinie überschritten hat etc. In diesen Fällen hat der Schiedsrichter einen Interpretationsspielraum, der Ausdruck einer nicht vollständigen Objektivität ist.
Immerhin bestünde die Möglichkeit, durch Video-Aufzeichnungen diesen Interpretationsspielraum einzuschränken, da die betreffende Spielsituation beliebig oft reproduziert werden könnte. Bestrebungen in dieser Richtung hatten bisher allerdings nur geringen Erfolg, wohl aus zwei Gründen:

a) stünde der notwendige technische Aufwand vielleicht in keinem Verhältnis zum Ergebnis,
b) würden solche Entscheidungen anhand von Video-Aufzeichnungen möglicherweise die Entscheidung des Schiedsgerichts sehr lange hinauszögern (vor allem, wenn auch die Aufnahmen kein eindeutiges Urteil zulassen), und damit den Wettkampf sehr lange unterbrechen,

c) andererseits können viele Entscheidungen auch nicht auf später verschoben werden, da z. B. ein Tennisspieler aus taktischen Gründen unmittelbar den Punktestand kennen muß.

Zum anderen interessiert uns in der Schule nicht nur die Leistung einer Mannschaft, wie sie etwa in der Torzahl zum Ausdruck kommt, sondern wir möchten im Hinblick auf eine Zensur wissen, was der einzelne im Spiel leistet. Nun hängt die Leistung des einzelnen in hohem Maße von seinen Mitspielern ab, und es gibt kein Mittel, den Einfluß der Interaktion auf die individuelle Leistung zu objektivieren. Der Lehrer muß sich (ähnlich wie bei den Kunst-Sportarten) immer auf seinen subjektiven Eindruck verlassen.
Der Versuch, den Einfluß der Interaktion durch Isolierung einzelner Spielelemente auszuschalten, indem man einen einzelnen Schüler nur dribbeln oder auf den Korb werfen läßt (und die Anzahl der erzielten Körbe als Meßwert für das komplexe Merkmal „Basketball-Spielen" nimmt), führt zwar zu objektiveren Meßwerten, wirft dafür aber andere Probleme auf: durch Summierung von Einzelteilen ist nur unzureichend die Gesamtheit des Spiels repräsentiert (gemäß dem Grundsatz der Gestaltpsychologie „Das Ganze ist mehr als die Summe seiner Teile"): ein Schüler, der alleine sehr gut mit dem Ball dribbelt, kann in der komplexen Situation durchaus ein schlechter Spieler sein.

Neben diesen technischen sind auch taktische Fertigkeiten ein ganz wesentliches Merkmal der Leistung eines Spielers:
Während die technischen Fertigkeiten noch relativ leicht und auch objektiv quantifizierbar sind (s. o., vgl. auch hierzu die KMK-Vereinbarungen), ist dies bei taktischen Leistungen nur sehr schwer und mit geringer Objektivität möglich. (Ganz abgesehen davon, daß diese beiden Spielleistungen auf unterschiedlichem Skalenniveau gemessen und dann miteinander verrechnet werden müssen.)
Versuche, taktisches Spielverhalten möglichst objektiv zu messen, haben notgedrungen nur beschränkten Erfolg. So sind etwa die „Analysebogen", wie sie in den KMK-Vereinbarungen (Seite 80) empfohlen werden, zeitlich allzu aufwendig und schließen subjektive Fehler keineswegs aus. (Darüber darf der scheinbar exakte, aber statistisch und logisch in keiner Weise begründete Verrechnungsmodus der Beobachtungsdaten nicht hinwegtäuschen; auch die schönste Verrechnung kann aus subjektiven Daten keine objektive Leistungsmessung machen.): Wir sehen einen Schüler A, der in recht günstiger Wurfsituation vor dem Korb steht; er wirft jedoch nicht, sondern spielt Schüler B an, der aus seiner Sicht in einer günstigeren Wurfposition steht. Schüler B wirft daneben, er stand doch nicht so günstig, wie A angenommen hatte. A hat taktisch klug gehandelt, sich aber in den Voraussetzungen getäuscht und objektiv das Gegenteil von dem erreicht, was er wollte. Wie bewerten wir das? Von der Intention her positiv, von der Effektivität her negativ. Was bewerten wir höher? Zwei verschiedene Lehrer werden hier möglicherweise zu zwei völlig unterschiedlichen Bewertungen kommen, d. h. nach der o. a. Definition: *Die Leistungsmessung im taktischen Bereich und damit auch im gesamten Spielbereich hat eine geringe Objektivität.*

U. a. aufgrund dieser Probleme nimmt eine Expertengruppe zu den KMK-Beschlüssen folgendermaßen Stellung:

a) „Es gibt kein standardisiertes Verfahren der Messung und Beurteilung von Leistungen im Sportspiel, durch das Leistungsnachweise im Spiel in eine Notenskala eindeutig überführt werden können. (Gemeint ist hier wohl: objektiviert werden können; Anm. d. Verf.)
b) Solche Standardisierung von Spielleistungen ist unter Rückgriff auf spieltheoretische Erkenntnisse weder möglich noch prinzipiell wünschenswert.
c) Ein objektives und standardisiertes Instrument zur pädagogischen Bewertung von Spielleistungen wird für unmöglich gehalten" (Zft f. Sportpäd. 3/77, 358 f.). In dem Versuch, „eine demnach vertretbare Lösung zu finden", schlagen die Autoren u. a. vor: „Das Gesamturteil des Lehrers muß auf dem Hintergrund der verfolgten didaktischen Konzeption des durchgeführten Sportunterrichts gefällt werden" (Seite 359). Es ist evident, daß der „Hintergrund der verfolgten didaktischen Konzeption" nicht objektivierbar ist im Sinne o. a. Definition und deshalb auch keine Objektivität der Leistungsfeststellung im Bereich des Spielens ermöglicht.

3.2 RELIABILITÄT

„Unter Reliabilität eines Testes versteht man den Grad der Genauigkeit, mit dem er ein bestimmtes Persönlichkeits- oder Verhaltensmerkmal mißt, gleichgültig, ob er dieses Merkmal auch zu messen beansprucht (welche Frage ein Problem der Validität ist)" (Lienert 1967, 14).

Das Problem der Reliabilität ist eine Frage nach der *Zuverlässigkeit* des Meßinstruments, mit dem man ein Merkmal erfassen will. Dabei geht man aus von der Annahme, daß das Merkmal konstant ist, und daß Schwankungen zwischen verschiedenen Messungen eben auf zufälligen Einflußfaktoren beruhen, die das Meßergebnis verfälschen, d. h. die zugrundeliegende Leistungsfähigkeit nicht angemessen im Ergebnis zum Ausdruck kommen lassen.

Auf den Bereich der sportlichen Leistungsmessung übersetzt würde das bedeuten: unsere Leistungsmessung ist dann reliabel (= zuverlässig), wenn wir bei verschiedenen Messungen zu verschiedenen Zeitpunkten dieselben Ergebnisse erhalten.

In der Entwicklung z. B. psychologischer Tests versucht man, den Einfluß solcher zufälliger Variablen zu minimieren. *Eine* Möglichkeit zur Verbesserung der Reliabilität eines Tests liegt darin, die Anzahl der Testaufgaben zu erhöhen: die „Intelligenz" eines Probanden wird nicht mit Hilfe *einer* Rechenaufgabe ermittelt, da hier der Zufall — positiv wie negativ — eine zu große Rolle spielen würde, sondern man gibt ihm möglichst viele Aufgaben, die er zu lösen hat.

(Um wieviel die Reliabilität eines Tests steigt, wenn man ihn um x Aufgaben verlängert, läßt sich nach der Brown-Spearmanschen Formel berechnen. Vgl. Lienert 1967).

In einigen Bereichen des Sports geht man ähnlich vor. So haben z. B. die Werfer und Weitspringer in der Leichtathletik drei bzw. sechs Versuche, von denen nur

der beste in die Endwertung eingeht. Ähnlich haben Hoch- und Stabhochspringer auf jeder Höhe drei Versuche, etc.

Ballspielmannschaften tragen jeweils zwei Spiele gegeneinander aus, ein Hin- und ein Rückspiel, wodurch z. B. der Einfluß der Zufallsvariablen „auf eigenem Platz und vor heimischem Publikum zu spielen" neutralisiert werden soll.

Dieser Möglichkeit einer Verbesserung der Reliabilität durch „Verlängerung des Tests" ist allerdings in vielen Sportarten aus zeitökonomischen Gründen oder aus Rücksicht auf die körperliche Belastbarkeit der Sportler enge Grenzen gesetzt: Es verbietet sich von selbst, 10 000-m-Läufer innerhalb eines Wettkampfes mehrfach gegeneinander laufen zu lassen; ein Kunstturner wäre wahrscheinlich kräftemäßig, der Veranstalter eines Kunstturnwettkampfes sicherlich zeitlich überfordert, wenn jeder Turner jede Übung dreimal vorführen dürfte.

Nun wissen wir aber, daß gerade im Sport die persönliche Bestleistung — und gerade die interessiert uns ja im Hinblick auf die Note — nicht der jederzeit beliebig reproduzierbare Normalfall sondern die Ausnahme darstellt. Es ist ein Merkmal dieser Grenzleistung, daß sie mal gelingen, mal mißlingen kann, daß sie nicht völlig konstant ist.

Ein Schüler, der an einem Tag 12,50 m stößt, erreicht vielleicht eine Woche später nur 11,50 m. Die Ursachen für diese Leistungsschwankungen können einmal im Sportler selbst liegen: seine Technik ist noch nicht ausgefeilt, der Bewegungsablauf noch nicht automatisiert, er ist unkonzentriert oder übermüdet etc. D. h. das Merkmal „sportliche Leistung" kann aufgrund individueller Dispositionen recht instabil sein — und ein Hauptziel des Trainings ist die Erhöhung der Stabilität: der Sportler möchte möglichst konstant nahe seiner optimalen Leistung bleiben (die Reliabilität seiner Testleistungen erhöhen).

Daneben können auch äußere Faktoren die Konstanz der Leistungen beeinträchtigen: schlechtes Wetter, zu glatter Boden, Gegenwind, keine Möglichkeit zum Aufwärmen etc. D. h. die Leistung wird nicht immer unter den gleichen äußeren Umständen vollzogen. Für einen Wettkampf ist das nicht von unmittelbarer Bedeutung, wenn alle Teilnehmer gleichermaßen diesen negativen Bedingungen ausgesetzt sind, wohl aber, wenn die Leistungen gleichzeitig als Qualifikation z. B. für die Aufnahme in eine Mannschaft oder für die Zulassung zu einer Meisterschaft dienen sollen — im Hinblick auf unser Problem: wenn die Leistungsmessung Grundlage für eine Zensur sein soll, wenn also das Ergebnis über den unmittelbaren Ablauf hinaus („Wer ist Sieger des 100-m-Laufs?") Bedeutung und Konsequenzen hat („Welcher Schüler ist im allgemeinen der beste Läufer?").

In Spielen ist die Reliabilität der Leistungsmessung noch dadurch reduziert, daß — aufgrund der Komplexität des Geschehens — die Leistung des Schülers A sehr stark von der Leistung seiner Mitspieler und/oder Gegenspieler abhängt. Sehr deutlich sehen wir das z. B. im Tennis: gegen einen gleichwertigen oder etwas besseren Partner spielen wir besser als gegen einen deutlich unterlegenen. Es ist kaum möglich, die Größe dieser Abhängigkeit abzuschätzen und in der Messung (und später in der Zensur) entsprechend zu berücksichtigen.

Die Konstanz der Leistung (und damit die Reliabilität der Messung) nimmt zu

a) mit dem Trainingszustand des Sportlers bzw. der Automatisierung der betreffenden Bewegungsabläufe,
b) mit der Standardisierung der äußeren Bedingungen (was z. B. für einen 5 000-m-Läufer leichter ist als für einen Ski-Langläufer),
c) mit der Entfernung von der Grenzleistung (d. h. ein Läufer, der — wenn auch selten — in der Lage ist, 11,0 sec zu erreichen, wird mit großer Wahrscheinlichkeit auch bei schlechten inneren und äußeren Bedingungen 11,6 sec erzielen, mit einiger Sicherheit aber 12,2 sec),
d) je weniger der Sportler im Leistungsvollzug auf andere Personen (z. B. im Spiel) angewiesen ist.

Nun kann das Problem der abnehmenden Konstanz der Leistung bei Annäherung an die Leistungsgrenze bei Schülern dadurch gemildert werden, daß wir die Leistungsmessung mehrfach wiederholen: er läuft mehrmals 100 m, turnt mehrmals seine Reckübung etc., wobei jeweils nur die beste Leistung als Meßwert herangezogen wird. Dieses Vorgehen verbietet sich bei den meisten Testverfahren in anderen Leistungsbereichen: wenn derselbe Intelligenztest mit kurzem zeitlichen Abstand wiederholt wird, wird der Proband beim zweiten Mal wahrscheinlich ein besseres Ergebnis erzielen, da er jetzt mit den Testaufgaben bereits vertraut ist und sich an einen Teil der Lösungswerte erinnert, aber nicht, weil er inzwischen intelligenter geworden ist. Ein ähnlicher Wiederholungseffekt würde sich einstellen, wenn wir z. B. dieselbe Klassenarbeit nochmals schreiben ließen; die zweite Arbeit würde wahrscheinlich besser ausfallen, obwohl die Schüler in der Zwischenzeit sicherlich keine besseren Mathematiker geworden sind.
Die Messung durch Tests unterscheidet sich von der Messung im Sport dadurch, daß ein Test immer nur eine (möglichst repräsentative) Verhaltens- oder Leistungsstichprobe erfaßt: eine Aufgabe im Intelligenztest soll auf das viel komplexere und umfassendere Merkmal „Intelligenz" schließen lassen, die Mathematikaufgabe auf „mathematisches Verständnis", während im 100-m-Lauf eben diese Leistung erfaßt werden soll, sie repräsentiert sich selbst. Deshalb kann der 100-m-Lauf auch beliebig oft wiederholt werden, ohne daß dadurch die Leistung verfälscht würde.
Die Wiederholung eines Intelligenztests ändert nicht das zu messende Merkmal, die Intelligenz, (der Proband ist beim 2. Mal nicht „intelligenter"), sondern ändert das Meßinstrument, den Test; (er prüft z. B. beim 2. Mal das Erinnerungsvermögen des Prüflings).
Die Wiederholung eines 100-m-Laufs hingegen ändert durch die Trainingswirkung das zu messende Merkmal positiv (der Schüler läuft beim 2. Mal ggf. tatsächlich schneller), während die Messung selbst unbeeinflußt bleibt.

Von Messungen in anderen Bereichen unterscheidet sich die sportliche Messung auch noch dadurch, daß positive Zufälle mit wenigen Ausnahmen ausgeschlossen sind.

Wenn ein Proband in einem Test mit Mehrfach-Wahl-Antworten eine Antwort weiß, kreuzt er die richtige Antwort an, weiß er sie nicht, kann er durch Raten rein durch Zufall die richtige Lösung finden, d. h. normalerweise wird durch zufällig richtige Lösungen das Testergebnis verbessert, während zufällig falsche Lösungen weit seltener auftreten.
Im Sport ist es umgekehrt: niemand kann „zufällig" schneller 100 m laufen oder hochspringen, als er aufgrund seiner physiologischen und technischen Bedingungen in der Lage ist (er kann allenfalls zufällig den Basketballkorb treffen), wohl aber kann er zufällig (etwa aufgrund äußerer oder innerer Bedingungen) *weniger* leisten als er eigentlich könnte.
Während wir also bei vielen Testverfahren eher ein *zufällig besseres* Abschneiden des Probanden erwarten können, müssen wir beim Sport eher mit einem *zufällig schlechteren* Abschneiden rechnen.
Sowohl im Bereich des Wettkampfsports als auch (normalerweise) im pädagogischen Bereich wird dieser Überlegung dadurch Rechnung getragen, daß von mehreren Leistungen jeweils die beste entscheidet.
(Dieser Unterschied zwischen kognitiver und motorischer Leistung im Hinblick auf zufällig richtige Lösungen ist m. W. bei der Entwicklung motorischer Tests noch nicht genügend berücksichtigt und die entsprechenden testtheoretischen Probleme sind nicht gelöst.)

Darüber hinaus ist das Meßinstrument selbst in vielen Sportarten nur bedingt reliabel, zuverlässig. Wie die Objektivität, so ist auch die Reliabilität im Bereich der C-G-S-Sportarten am größten: ein Bandmaß, eine Stoppuhr, eine Waage zeigen (sofern sie in Ordnung sind) auch bei wiederholten Messungen einer Leistung konstant denselben Wert an (wenn man davon absieht, daß z. B. ein Bandmaß eine gewisse Elastizität besitzt oder die Länge geringfügig von der Temperatur abhängt; dieser Fehler kann jedoch bei großen Sportveranstaltungen durch optische Meßverfahren ausgeschlossen werden.).
In den Wertungssportarten (Turnen etc.) ist die Reliabilität des Meßinstruments erheblich eingeschränkt: das „Meßinstrument" (beim 100-m-Lauf die Stoppuhr) ist hier der Kampfrichter, dessen Urteil bei einer wiederholten Bewertung desselben Bewegungsablaufes (z. B. an einer Film- oder Video-Aufzeichnung) erheblich schwanken kann. Er funktioniert nicht wie eine Uhr, seine Wahrnehmung wird von zahlreichen und z. T. unkontrollierbaren Faktoren beeinflußt (vgl. Ingenkamp 1973, Kleber).
Wir sehen also: Die *Reliabilität* der Leistungsmessung ist zum einen durch die Abhängigkeit der Leistung von äußeren Einflüssen und die möglicherweise geringe Konstanz des Merkmals, anderseits besonders in den Kunstsportarten durch die Wahrnehmungsfehler und Urteilsschwankungen bei den Punktrichtern beeinträchtigt.
Ferner gilt auch für den Bereich des Sports das Gesetz aus der Testtheorie: Die Reliabilität einer Leistungsmessung kann nicht größer sein als ihre Objektivität (vgl. Lienert 1967, 19—21).

3.3 VALIDITÄT

„*Validität*" ist ein Begriff aus der Testtheorie und bedeutet nach einer bekannten Definition von LIENERT: „Die Validität eines Testes gibt den Grad der Genauigkeit an, mit dem dieser Test dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das (die) er messen soll oder zu messen vorgibt, tatsächlich mißt. Ein Test ist demnach vollkommen valide, wenn seine Ergebnisse einen unmittelbaren und fehlerfreien Rückschluß auf den Ausprägungsgrad des zu erfassenden Persönlichkeits- oder Verhaltensmerkmals zulassen" (Seite 16). So wäre z. B. ein Test, der vorgibt, die Konzentrationsleistung zu prüfen, bei dem es aber sehr stark auf das Sprachverständnis ankommt, sehr wenig valide.

Die Frage nach der Validität der Leistungsmessung im Sport muß zumindest auf drei Ebenen beantwortet werden.

1. Durch die Wettkampfregeln einer bestimmten Disziplin wird die Validität eines Meßvorgangs (meist) eindeutig operational definiert: Die Leistung des Weitspringers ergibt sich aus dem auf 1 cm genau gemessenen Abstand zwischen Vorderkante des Sprungbalkens und hinterstem Eindruck des Springers. Wenn wir wissen wollen, wer der beste Weitspringer ist, so läßt der gemessene Wert „einen unmittelbaren und fehlerfreien Rückschluß" zu: Bester ist — den Wettkampf gewinnt — wer bei einer solchen Leistungsmessung den höchsten Wert erzielt —. Das erscheint uns trivial.

   In der Testtheorie würde man von „inhaltlicher Validität" sprechen: „Der Test selbst stellt das optimale Kriterium für das Persönlichkeitsmerkmal dar" (LIENERT 1967, 16).

2. Nicht so einfach stellt sich die Frage nach der Validität, wer — den Wettkampfregeln entsprechend — der beste Weitspringer ist, sondern wer am weitesten springen kann, wer die beste Sprungkraft besitzt.
   Stellen wir uns vor:

   A springt genau an der Balkenvorderkante ab und erzielt 6,00 m.

   B springt 10 cm vor der Vorderkante ab (er „verschenkt") und erzielt beim regelentsprechenden Meßvorgang 5,95 m — wird also schlechter eingestuft als A, obwohl er objektiv 5 cm weiter gesprungen ist.

   Die wettkampfmäßige Weitsprungleistung hängt u. a. sehr stark von der Anlaufgenauigkeit ab, die oft erst durch einen mühsamen Trainingsprozeß erworben werden muß. Das gemessene Merkmal „Weitsprungleistung" läßt also keinen unmittelbaren Schluß auf eine zugrunde liegende Fähigkeit „Sprungkraft" zu.

   Wenn es nun Anliegen der Schule ist, in den unteren Jahrgängen in erster Linie grundlegende Fähigkeiten zu schulen (ggf. unter Verzicht auf offizielle Wettkampfformen), so ist es nur folgerichtig, daß bei den Bundesjugendspielen bis zu einer bestimmten Altersstufe aus einer Absprungzone gesprungen werden darf, so daß die Anlaufgenauigkeit eine geringere Rolle spielt und der Meßwert valider die Sprungleistung wiedergibt.

Das Problem, wie weit auf dieser 2. Ebene von der gemessenen Leistung unmittelbar auf eine entsprechende zugrunde liegende Leistungsfähigkeit geschlossen werden kann, wird in den Wertungssportarten besonders deutlich, weil hier ganz offen zwei verschiedene Merkmale (Schwierigkeit der Übung und Präzision der Ausführung) erfaßt werden müssen.

Veranschaulichen wir uns das an folgendem Beispiel:

A springt in einem Wettkampf/einer Prüfungssituation einen einfachen Salto, obwohl er im Training auch den Doppelsalto gut beherrscht — er will kein Risiko eingehen. Er erhält die Haltungsnote 5. Der Schwierigkeitsgrad für den einfachen Salto gehockt beträgt 1,3, er erhält also insgesamt 6,5 Punkte.

B springt ebenso gut wie A, traut sich aber mehr zu und zeigt einen Doppelsalto. Schwierigkeitsgrad 2,0, Haltungsnote (ebenso wie bei A) 5 Punkte — B erhält insgesamt 10 Punkte.

D. h. die Wertungen 6,5 und 10 Punkte geben an, was die gezeigten Sprünge „wert" sind, sie lassen aber nur bedingt einen Rückschluß auf die (motorische!) Leistungsfähigkeit der Springer zu, da die größere Risikofreude des Springers B sicherlich kein motorisches Merkmal darstellt, sondern anderen Persönlichkeitsbereichen zuzuordnen ist.

Wieder anders stellt sich das Problem bei den Spielen dar: Wenn ein Tennisspieler A den Spieler B mit 6:2, 6:2 besiegt — und gemäß den Spielregeln drückt sich bei einem Wettspiel die Leistung eben in den Punkten aus — so wird in diesem „Meßwert" (und das Zählen der Punkte ist eine Art des Messens, vgl. Kapitel 2.2.1) die Leistung *beider* Spieler ausgedrückt; d. h. wir können nur dann auf die Leistung des Spielers A schließen, wenn wir die (sonst übliche) Leistung von B kennen. Der Meßwert 6:2, 6:2 ist keine individuelle Meßzahl eines der beiden Spieler.

Hier kann sich allerdings ein Mißverständnis einschleichen: die „sonst übliche" Leistung von B können wir auch immer nur aufgrund der Vergleiche mit anderen kennen. Zu diesem Zweck stellt man z. B. Ranglisten auf, die nach der Methode „jeder gegen jeden" (z. B. im Fußball) im Rangplatz die Leistungsstärke des einzelnen Spielers oder einer Mannschaft ausdrücken soll.

In den Spielen (ebenso in den Kampfsportarten) ist also aufgrund der Abhängigkeit der gemessenen Leistung von der Leistung des Partners oder Gegners kein unmittelbarer Rückschluß auf die individuelle Leistung des einzelnen möglich.

In ähnlicher Weise hat sich diese Abhängigkeit auf die Reliabilität ausgewirkt; wie für die Reliabilität gilt für die Validität: Die Validität einer Messung kann nicht größer sein als deren Reliabilität.

3. Auf der 3. Ebene wird die Frage nach der Validität der Leistungsmessung im *Sportunterricht* noch dadurch kompliziert, daß in unseren Zeugnissen eine Zensur für „Sport" (oder Leibeserziehung, Leibesübungen etc.) erteilt wird; das ist ein recht komplexer Bereich, dessen Definition — und das ist die

Voraussetzung für eine valide Messung — recht problematisch ist. Die Zensur beruht normalerweise (analog einem Test) auf der Lösung verschiedener Aufgaben, auf Leistungsmessungen im Turnen, in den Spielen, im Schwimmen, der Leichtathletik etc. Die Summe dieser Leistungsmessungen ergibt eine „Rohpunktsumme"; die in eine Zensur transformiert wird. (Wir wollen uns an dieser Stelle auf den sportmotorischen Bereich beschränken; es ist üblich, daß noch andere Merkmale wie Begabung, Leistungswille etc. in der Note berücksichtigt werden — das wird im nächsten Kapitel besprochen.)

Ob nun diese Leistungserhebungen, die wir im Verlaufe eines Schulhalbjahres vornehmen, repräsentativ für den „Sport" sind, d. h. ob sie eine valide Aussage über die sportlichen Fähigkeiten des Schülers machen, hängt weitgehend von der Definition dieses Leistungsbereiches ab, konkret: von den Sportarten, die wir unterrichten und prüfen, und von der Entscheidung, ob wir nur dasjenige bewerten, was wir in der Schule unterrichten oder ob wir auch außerschulische Leistungen berücksichtigen wollen.

Aus der Testpsychologie wissen wir (Meili 1964, Guilford 1958), daß motorische Aufgaben nur sehr niedrig miteinander korrelieren; wer z. B. gut balancieren kann, ist möglicherweise in der Reaktionsgeschwindigkeit oder bei Schnelligkeitsaufgaben schlecht (Vgl. auch Volkamer 1971).

Nach unseren Erfahrungen gilt das nicht nur für relativ isolierte und eng begrenzte Testaufgaben, sondern auch für die verschiedenen, z. T. recht komplexen sportlichen Disziplinen: ein guter Turner ist vielleicht ein mittelmäßiger Spieler und ein miserabler Schwimmer. (Selbstverständlich haben wir auch Schüler, die in allen Disziplinen etwa gleich gut oder etwa gleich schlecht sind, das dürfte sogar in der Mehrzahl der Fälle zutreffen, und das würde auch in einem positiven Korrelationskoeffizienten zum Ausdruck kommen.) Die Beobachtung, daß das Leistungsniveau eines Schülers in verschiedenen Disziplinen verschieden hoch sein kann, hat ja auch zu dem Vorschlag geführt, nicht *eine* Zensur im Sport zu erteilen, sondern für jede im Unterricht eingeführte Disziplin eine gesonderte Note. Dieser Vorschlag hat sich nicht durchgesetzt, u. a. wohl deshalb, weil diese Art der Zensierung wesentlich — auch äußerlich optisch im Zeugnis — von der der anderen Fächer abweichen würde.

Nehmen wir nun an, Lehrer A will in der Zensur eine möglichst umfassende Aussage über die sportlichen Fähigkeiten und Fertigkeiten seiner Schüler machen. Konsequenterweise müßte er dann auch Leistungen außerhalb der Schule und in Sportarten, die er gar nicht unterrichtet, berücksichtigen (z. B. Tennis, Skilauf, Bergsteigen etc.). Die meisten Lehrer werden das — mit individuell sehr verschiedener Gewichtung — auch tun.

Eine Befragung von 137 Sporstudenten hat ergeben, daß 39% auch außerschulische Leistungen berücksichtigen wollen, wobei die Gewichtung (zu wieviel Prozent solche Leistungen maximal zu Buche schlagen sollten) zwischen 5% und 20% schwankte. Der Wert von 39% deckt sich genau mit den Befragungsergebnissen, die Kröner (1976) an 50 Sportlehrern gewonnen hat.

So plausibel uns die Berücksichtigung außerschulischer Leistungen auch erscheint, dieses Vorgehen ist unter drei Gesichtspunkten nicht unproblematisch:

1. Darf oder soll die Schule Leistungsbereiche zensieren, d. h. in einem schulischen Dokument zum Ausdruck bringen, die nicht in der Schule selbst unterrichtet werden? Die Antwort hierauf hängt davon ab, welche Aufgabe der Zensur zugesprochen wird. So ist es z. B. unter dem Aspekt des feedback für Schüler und Lehrer hinsichtlich der Annäherung des Schülers an ein vorgesehenes Lernziel (vgl. Kapitel 8.4) unsinnig, Leistungen zu berücksichtigen, die nicht in den Lernzielen gefordert sind.

2. Häufig ist der Lehrer in den außerschulischen Sportarten zu wenig bewandert, um die Leistung des Schülers angemessen erfassen zu können. Er weiß z. B., daß Schüler F regelmäßig reitet — wie soll das seine Sportzensur beeinflussen, wenn er nicht weiß, wie gut er reitet? Vielleicht ist er ja — obwohl er es regelmäßig tut — ein miserabler Reiter. D. h. er wird die außerschulische sportliche Aktivität nur dann positiv ins Zeugnis eingehen lassen können, wenn er die Einstellung hat, daß jede sportliche Betätigung unabhängig von der Leistungshöhe sich positiv auf die Zensur auswirken sollten.

   Solange wir aber in den Zensuren Relationen zwischen den Schülern einer Klasse herstellen (vgl. Kapitel 8.1) führt diese an sich sehr positive und für den betroffenen Schüler erfreuliche Zensierungspraxis zu einer Ungerechtigkeit für diejenigen Schüler, die nicht außerhalb der Schule aktiv sind, da sie nun relativ schlechter zensiert werden: Darf die Schule einen Schüler bestrafen, daß er nur die Anforderungen der Schule erfüllt und keine darüber hinaus gehenden Aktivitäten entwickelt? Das erschiene absurd.

3. Ein weiteres Problem dürfte sich dann ergeben, wenn der Schüler sich in seiner außerschulischen Aktivität in hohem Maße engagiert und hohe Leistungen bringt, am Schulsport aber keinerlei Interesse zeigt und weit Unterdurchschnittliches leistet.

   Hier hinge die Zensur also zum einen wiederum von der Gewichtung ab: wie stark sollten außerschulische Aktivitäten überhaupt die Zensur beeinflussen können? Zum anderen käme es in einem solchen Fall sicherlich zu persönlichen Problemen des Sportlehrers: er wird möglicherweise das Desinteresse an seinem Unterricht bei offenbar guter Leistungsfähigkeit als persönliche Ablehnung empfinden, als Ablehnung seines Unterrichts, und diese persönliche Betroffenheit wird die Objektivität seines Urteils sicherlich beeinträchtigen.

Haben wir bis hierhin die Probleme diskutiert, die sich aus einer weiten Definition des Leistungsbereichs „Sport" ergeben, betrachten wir nun eine enge Definition; Lehrer B geht möglicherweise von der Einstellung aus, daß die Schule nur dasjenige beurteilen dürfe, was im Unterricht durchgenommen wird; Aktivitäten außerhalb der Schule seien zwar erfreulich und wünschenswert, seien aber letztlich Privatvergnügen des Schülers.

Aus dieser Einstellung ergeben sich folgende Probleme:

Wir wissen, daß die Leistungen in verschiedenen Disziplinen z. T. sehr niedrig korrelieren. Ob ein Schüler nun eine gute oder eine weniger gute Note erhält, hängt u. U. davon ab, welche Sportart bevorzugt vom Sportlehrer unterrichtet wird. Der Vereinsschwimmer, dessen Lehrer Turnspezialist ist und vorwiegend Turnen unterrichtet, wird gegenüber einem Schüler, der im Verein turnt, benachteiligt sein, obwohl Schwimmen ebenso wie Turnen zu den „klassischen" Schulsportarten gehören. D. h. eine enge Begrenzung der Zensierung auf diejenigen Bereiche, die in der Schule auch vermittelt werden, hat zur Folge, daß

a) die vom Lehrer bevorzugten Sportarten die Note sehr stark beeinflussen werden,
b) die Anzahl der in die Note eingehenden Leistungsbereiche sinkt, und damit sinkt die Validität der Zensur, d. h. sie läßt nur einen sehr begrenzten Rückschluß auf den Leistungsbereich „Sport" zu: ein Schüler mit einer schlechten Sportnote kann durchaus ein leistungsstarker Sportler in seiner Freizeit sein.

In der eng begrenzten Definition wird letztlich willkürlich das umfassende und sehr komplexe Leistungsgebiet Sport reduziert auf dasjenige, was die Schule (bzw. der jeweilige Lehrer) aus diesem Gebiet herausgreift. Die gemessenen Leistungen sind ggf. nicht mehr repräsentativ für den Sport, und die Leistungsmessung läßt keinen „unmittelbaren und fehlerfreien Rückschluß auf den Ausprägungsgrad des zu erfassenden Persönlichkeits- oder Verhaltensmerkmals" zu, d. h. sie ist nicht oder nur eingeschränkt valide.

Wir können zusammenfassen:

Die Validität der Leistungsmessung kann nicht größer sein als ihre Reliabilität und ihre Objektivität. Sie ist am höchsten dort, wo sie inhaltlich aufgrund der Wettkampfregeln definiert ist, und wird umso niedriger, je mehr wir aufgrund der Leistungsmessung auf zugrunde liegende Fähigkeiten (z. B. Sprungkraft) schließen oder gar eine generelle Aussage über „die" Sportlichkeit des Schülers machen wollen.

Mit der Diskussion der Validität berühren wir das didaktische Problem, daß wir definieren müssen, was Sport und speziell der Schulsport ist, und welche Aufgaben er hat, kurz: was ein „guter" Sportler ist.

Solange dieses Problem nicht gelöst ist, bleiben die Bemühungen um die Erhöhung der Objektivität und Reliabilität in der schulischen Leistungsmessung und Leistungsbewertung fruchtlos.

4. Untersuchung zur Objektivität und Reliabilität bei der Bewertung (Messen) von motorischen Leistungen

4.1 OBJEKTIVITÄT

„Eine Hilfe für leistungsschwächere (konstitutionell benachteiligte) Schüler bedeutet es, wenn wir neben der quantitativen Leistung, dort wo nach dem C-G-S-System gemessen wird, auch die qualitative, die eigentliche Lernleistung bewerten. Beispiel „Hochsprung": Gemessen wird bei jedem Schüler die gesprungene Höhe und zusätzlich eine Sprungtechnik gewertet, die über eine frei gewählte Höhe vorgeführt werden kann. In den angesprochenen Bereichen bzw. Disziplinen sollte die Bewertung von Bewegungsqualitäten, also von Lernleistungen, generell zumindest gleichrangig neben der quantitativen Leistungsmessung stehen" (GÜNZEL 1977, 101).

Um die Objektivität und Reliabilität der Bewertung (= Messung) solcher „qualitativen Lernleistungen" zu prüfen, wurde von uns folgender Versuch durchgeführt: 143 Sportstudenten der Universitäten Münster (94 Vpn) und Osnabrück (49 Vpn) sollten Bewegungsabläufe aus der Leichtathletik beurteilen. Dazu erhielten sie das Schema auf S. 42.

Es wurde ihnen ein Super-8-Film gezeigt, in dem ein Student in wechselnder Reihenfolge vier Hürdenläufe (jeweils über drei Hürden), zwei Speerwürfe, zwei Starts und vier Kugelstöße vorführte. Nach jeder Disziplin wurde der Film angehalten und die Studenten gebeten, ihre Wertung in den Bogen einzutragen. Danach wurde der nächste Bewegungsablauf gezeigt, bewertet etc., so daß am Ende jede Vp bei zwölf Bewegungsabläufen insgesamt 46 Einzelwertungen abgegeben hatte (bei 143 Vpn insgesamt 6578 Einzelwertungen).

Für diese Untersuchung hatten wir Disziplinen aus der Leichtathletik ausgewählt, weil hier die Einzelbewegung recht kurz und engumschrieben ist, und zudem ein recht genaues Bild von der Idealform vorliegen dürfte. Im Gegensatz zum Turnen wurde hier also weniger eine Beurteilung nach „schön — weniger schön" als vielmehr nach „richtig — weniger richtig" erwartet.

Ergebnisse

Als erstes wurden die Häufigkeiten für jeden Punktwert und jede Zelle ausgezählt; wir stellten fest, wie oft für die vorgeführten Bewegungsabläufe in den vorgegebenen Kategorien (also z. B. beim Hürdenlauf: 1. Höhe und Körperlage über die Hürde, 2. Körperlage bei der Landung etc.) jeweils die Wertungen 1, 2, 3, 4 oder 5 gegeben wurden. Wir erhielten also 46 derartige Häufigkeitsverteilungen.

Wir können hier nicht alle Verteilungen wiedergeben, sondern beschränken uns exemplarisch auf jeweils einen Bewegungsablauf aus jeder der vier Disziplinen.

Die Punkteverteilung dieser Beispiele unterscheidet sich nicht von den übrigen Bewertungen.
Gehen Sie bei der Wertung davon aus, daß es sich um einen Sportstudenten handelt, der seine Leichtathletikausbildung abgeschlossen hat.

Wertung

	sehr gut	noch gut	mittel-mäßig	eher schlecht	sehr schlecht
Punkte:	5	4	3	2	1

Bitte versuchen Sie, wirklich zu differenzieren und auch die Extremwerte zu benutzen, — die Bewegungsabläufe sind nicht fehlerfrei!

Hürdenlauf	Punkte				
Höhe und Körperlage über der Hürde					
Körperlage bei der Landung					
rhythmischer Ablauf					
Gesamteindruck					
Speerwurf					
Armführung					
Rhythmus der letzten 5 Schritte					
Spannung in der Abwurfphase					
Gesamteindruck					
Start					
Körperstellung bei „Auf die Plätze"					
Körperstellung bei „Fertig"					
Gesamteindruck					
Kugelstoß					
Ausgangslage					
Angleiten					
Spannung in der Abwurfphase					
Gesamteindruck					

Beurteilungsbogen: Bewegungsabläufe aus der Leichathletik.

Die Prüfung auf Normalverteilung erfolgte über die Summe der quadrierten Werte für Schiefe und Exzeß.

	Punkte					Normalverteilung
	1	2	3	4	5	
1. Kategorie	0	22	32	64	25	—
2. Kategorie	1	21	53	59	9	+
3. Kategorie	1	12	30	69	31	—
4. Kategorie	0	8	47	72	16	+

Tabelle 1 a Hürdenlauf

	Punkte					Normalverteilung
	1	2	3	4	5	
1. Kategorie	14	31	42	48	8	—
2. Kategorie	16	47	47	26	7	+
3. Kategorie	16	44	52	27	4	+
4. Kategorie	9	40	59	32	3	+

Tabelle 1 b Kugelstoß

	Punkte					Normalverteilung
	1	2	3	4	5	
1. Kategorie	3	36	46	50	8	—
2. Kategorie	33	44	36	29	1	—
3. Kategorie	43	62	25	11	2	—
4. Kategorie	16	51	58	17	1	+

Tabelle 1 c Speerwurf

	Punkte					Normalverteilung
	1	2	3	4	5	
1. Kategorie	5	23	45	49	21	—
2. Kategorie	15	49	41	28	10	—
3. Kategorie	8	44	54	31	6	+

Tabelle 1 d Start

Wir sehen also, daß die Einschätzung des Bewegungsablaufes durch die Beobachter *über die ganze Skalenbreite streut*, d. h. derselbe Ablauf wird von einem Teil als „sehr gut" (5 Punkte) wahrgenommen, während ihn andere als „sehr schlecht" (1 Punkt) beurteilen. Daß die mittleren Werte am häufigsten auftreten, dürfte zum einen an der tatsächlichen Qualität der Demonstration liegen, zum anderen an der (bei solchen Beurteilungen immer zu beobachtenden) Tendenz, Extremwerte zu meiden.
(Von den insgesamt 46 Kategorien sind (bei einem p-Wert = 0.05) 24 normalverteilt und 22 nicht normalverteilt bewertet worden.)
Unter den Bedingungen dieses Experiments können wir also schließen, daß die *Übereinstimmung zwischen verschiedenen Beurteilern sehr gering ist, daß es weniger auf die gezeigte Leistung als vielmehr auf den Beobachter ankommt, welchen Meßwert dem Bewegungsablauf zugesprochen wird.*

Demnach hat eine derartige Bewertung eine *sehr niedrige Objektivität.*

Um unsere Ergebnisse zu kontrollieren, wurden die Vpn in drei Gruppen eingeteilt:

1. Gruppe: Studenten, die noch nicht an der Leichtathletikausbildung teilgenommen hatten;
2. Gruppe: Studenten, die mitten in der Leichtathletikausbildung standen;
3. Gruppe: Studenten, die ihre Leichtathletikausbildung bereits abgeschlossen hatten.

Mit Hilfe der Varianzanalyse wurde geprüft, ob die drei Gruppen sich hinsichtlich ihrer durchschnittlichen Beurteilung unterscheiden.
In 29 von 46 Kategorien zeigten sich keine Unterschiede, d. h. der Ausbildungsstand hatte keinen signifikanten Einfluß auf die Beurteilung. In 17 Fällen traten signifikante oder sehr signifikante Unterschiede auf, und zwar mit der (sehr signifikanten Tendenz), *daß die Bewertungen mit dem Ausbildungsstand sinken:* je weiter die Studenten in ihrer Leichtathletikausbildung fortgeschritten sind, desto geringere Punktzahlen vergeben, desto strenger bewerten sie.
Es ist sinnvoll anzunehmen, daß eine intensive Leichtathletikausbildung einerseits die Vorstellung vom idealen Bewegungsablauf festigt, und daß andererseits durch das längere Sportstudium allgemein die Fähigkeit zu differenzierterer Bewegungswahrnehmung gesteigert wird. Demnach wäre also zu erwarten, daß die älteren Semester stärker in ihren Beurteilungen übereinstimmen.
Um diese Annahme zu prüfen, wurde nach den drei Gruppen getrennt für jede Zeile der Variabilitätskoeffizient berechnet. (Der Variabilitätskoeffizient ermöglicht es, bei Gruppen mit verschiedenen Mittelwerten festzustellen, ob das gemessene Merkmal in Gruppe A stärker streut als in Gruppe B. Vgl. dazu Clauss / Ebner 1972, 89). Dabei zeigte sich, daß die Gruppe, die in der Leichtathletikausbildung am weitesten fortgeschritten war, in ihrer Bewertung signifikant am häufigsten den *größten* Variabilitätskoeffizienten zeigte. Der mittlere Variabilitätskoeffizient betrug für die drei Gruppen:

a) Vpn ohne Leichtathletikausbildung V = 30.03
b) Vpn mit einem Teil der Ausbildung V = 29.00
c) Vpn mit abgeschlossener Ausbildung V = 33.21

Also auch ein Vergleich der über alle 46 Beurteilungen gemittelten Variabilitätskoeffizienten weist die 3. Gruppe als diejenige aus, die in ihrer Beurteilung am uneinheitlichsten war.
Unsere oben formulierte Erwartung wurde nicht bestätigt: *die längere Ausbildung führte also nicht zu einer Vereinheitlichung der Bewertungen.*
Wir können auch formulieren: *Durch längere Ausbildung steigt nicht die Sicherheit des Urteils, allenfalls die Selbstsicherheit der Beurteiler.* (Und es ist nicht zu erwarten, daß sich diese Entwicklung durch Berufseintritt ändert.)
Wir finden also für unseren Bereich die Vermutung Ingenkamps (1972) bestätigt, „daß erfahrene Praktiker und Experten kein zutreffenderes Eindrucksurteil über Mitmenschen haben als Neulinge" (Seite 16).
Nun lassen sich die Ergebnisse dieses Versuchs nicht ohne weiteres auf die Situation des Lehrers in der Schule verallgemeinern: der Lehrer sieht seine Schüler über längere Zeit in vielen verschiedenen Situationen und Disziplinen und nicht nur in einer „Momentaufnahme" wie hier im Film, d. h. die umfassendere Information könnte zu einer größeren Objektivität führen — anderseits werden vermutlich andere Faktoren wie Leistungswille, soziales Verhalten, Diszipliniertheit des Schülers etc. wiederum seine Wahrnehmung negativ beeinflussen (vgl. dazu Punkt 5.2). Ferner wird ein Lehrer, der gewohnt ist, immer allein zu unterrichten, mit der Zeit ein eigenes und — weil es kaum noch in Frage gestellt und korrigiert wird — recht rigides Wahrnehmungs- und Bewertungssystem entwickeln, was vielleicht auf die Dauer die Konstanz aber nicht die Objektivität seines Urteils erhöht. (Objektiv ist ein Meßverfahren, wenn die Ergebnisse unabhängig vom Untersucher sind.)

4.2 RELIABILITÄT

Nachdem wir also festgestellt hatten, daß die *Objektivität* unserer Beurteiler recht gering war, wollten wir prüfen, wie groß die *Reliabilität* ihres Urteils war, d. h. ob sie bei wiederholter Beurteilung desselben Bewegungsablaufs zu identischen Werten kommen würden (vgl. dazu auch Ingenkamp 1972).
Bei der Aufnahme der Bewegungsabläufe hatten wir zwei Kameras eingesetzt, so daß wir also von jedem Bewegungsablauf jeweils zwei Aufnahmen besaßen, die sich nur (z. T. sehr geringfügig) durch den Standort der Kamera oder durch die Laufgeschwindigkeit (18 oder 24 Bilder/sec) unterschieden. So waren in dem den Sportstudenten vorgeführten Filmstreifen jeweils zwei Bewegungsabläufe identisch (also z. B. der 1. und 3., 2. und 4. Hürdenlauf etc.), ohne daß die Vpn das wußten.

Die Wertungspunkte wurden unter zwei verschiedenen Gesichtspunkten ausgewertet:

1. Zuerst wurde der Mittelwert für jede Zelle errechnet, also z. B. für den 1. Hürdenlauf in der Kategorie „Höhe und Körperlage . . ." etc. Danach wurden die Mittelwerte der identischen Bewegungsabläufe miteinander verglichen und auf Signifikanz der Unterschiede geprüft.
   Von den 23 Mittelwertpaaren wichen nur sechs (!) nicht signifikant voneinander ab (p = 0.05), während 17 signifikante oder sogar sehr signifikante Unterschiede aufwiesen. Die größten Differenzen traten beim Speerwurf (2. Kategorie) und beim Start (1. Kategorie) auf: Beim Speerwurf betrug die mittlere Wertung im 1. Durchgang 3.19 Punkte, im 2. Durchgang 2.44 Punkte (Differenz = 0.75); beim Start (1. Kategorie) wurden beim 1. Durchgang durchschnittlich 2.57 Punkte, beim 2. Durchgang 3.40 Punkte gegeben (Differenz = 0.83).

Daraus läßt sich schließen:

Bei der wiederholten Bewertung desselben Bewegungsablaufes nach qualitativen Merkmalen kommen die Beurteiler in den meisten Fällen zu *signifikant unterschiedlichen Ergebnissen — der Meßvorgang besitzt geringe Reliabilität.*
Das entspricht den Befunden von Eells (1972), der denselben Aufsatz von denselben Lehrern hat mehrfach wiederholt zensieren lassen: „Die Variabilität des menschlichen Urteils beim selben Individuum ist etwa vergleichbar der Variabilität zwischen zwei Individuen" (In: Ingenkamp 1972, 122).
(Es ist nicht zu vermuten, daß allein die unterschiedliche Kameraeinstellung für diese Unterschiede verantwortlich gemacht werden kann. Wenn wir diesen Einfluß jedoch annehmen, so bedeutet das für die Schulpraxis: Die Note des Schülers hängt davon ab, ob ihn der Lehrer von links oder von rechts sieht. Vielleicht nicht zu unrecht bezeichnen die Schüler ihre Zeugnisse voller Ironie als „legales Glücksspiel".)

2. Mit Hilfe der Varianzanalyse wurde sodann geprüft, ob eine Wechselwirkung besteht zwischen dem Ausbildungsstand und der Bewertung im 1. und 2. Versuch.

   Am Beispiel des Hürdenlaufs (1. Kategorie):

	1. Wertung ($\bar{x}$)	2. Wertung ($\bar{x}$)	Differenz
1. Gruppe	3.48	3.35	0.13
2. Gruppe	3.96	3.30	0.66
3. Gruppe	3.79	3.05	0.74

Tabelle 2 Wechselwirkungen zwischen Wertung und Ausbildungsstand

   In diesem Beispiel unterscheiden sich die drei Gruppen sehr signifikant: Die Studenten, die ihre Leichtathletikausbildung bereits abgeschlossen hatten, zeigten die größte Differenz zwischen 1. und 2. Wertung. Das würde bedeuten, daß die längere Ausbildung *nicht* zu einer größeren Konstanz des Urteils führt (s. o.).

Die Tendenz ging allerdings dahin, daß in den meisten Fällen *keine signifikanten Unterschiede* auftraten (16 Wertungen nicht signifikant, 7 Wertungen signifikant oder sehr signifikant), d. h. die *Ausbildung hat in den meisten Fällen gar keinen Einfluß auf die Reliabilität der Wertung.* Das bestätigt das Untersuchungsergebnis im Hinblick auf die Objektivität (vgl. Kap. 4.1).

Anschließend wurde geprüft, wie stark jeweils die individuellen Wertungen im 1. und 2. Fall voneinander abwichen. Wir stellten für jede Vp die Differenz zwischen 1. und 2. Wertung (in absoluten Zahlen) fest und zählten dann aus, wie oft identische Werte, eine Differenz von 1, 2, 3 oder 4 Punkten auftrat.
Für die 143 Vpn ergaben sich bei den 23 Bewertungspaaren 3.289 Differenzen, die sich wie folgt verteilten:

Differenz	0 Punkte	=	1335 mal (40.58%)
	1 Punkt	=	1397 mal (42.47%)
	2 Punkte	=	454 mal (13.80%)
	3 Punkte	=	90 mal (2.73%)
	4 Punkte	=	13 mal (0.39%)

Tabelle 3 a Differenz zwischen 1. und 2. Wertung

Ca. 60% der Bewertungen fallen also unterschiedlich aus, bei 17% treten zwei oder mehr Punkte Unterschied auf, im Durchschnitt 0.8 Punkte.
Am deutlichsten waren die Differenzen zwischen 1. und 2. Wertung beim Start.

Differenz	0 Punkte	=	44 mal (30.8%)
	1 Punkt	=	52 mal (36.4%)
	2 Punkte	=	31 mal (21.7%)
	3 Punkte	=	11 mal (7.7%)
	4 Punkte	=	5 mal (3.5%)

Tabelle 3 b Differenz zwischen 1. und 2. Wertung beim Start

Das bedeutet in diesem Fall, daß nur rund 30% der Beurteiler in der 1. und 2. Bewertung identische Punkte vergaben, während immerhin noch 11% der Vpn denselben Ablauf mit drei oder vier Punkten Unterschied bewerteten; was also beim 1. Mal als „sehr gut" eingestuft wurde, erhielt beim 2. Mal den Wert „eher schlecht" oder sogar „sehr schlecht".
Abschließend stellen wir für jede Vp die durchschnittliche Differenz zwischen 1. und 2. Wertung über alle Beurteilungen fest und fanden dabei, daß es keine Vp gab, bei der alle Wertungen übereinstimmten.

Die 3 Gruppen zeigten folgende Gesamt-Mittelwerte für die Abweichungen zwischen 1. und 2. Beurteilung:

1. keine Ausbildung $\bar{x} = 0.80$
2. halbe Ausbildung $\bar{x} = 0.83$
3. ganze Ausbildung $\bar{x} = 0.83$

Auch dieses Ergebnis bestätigt unsere Beobachtung, daß die Dauer der Ausbildung offenbar keinen Einfluß auf die Sicherheit der Beurteilung hat.

Als Ergebnis unseres Film-Experiments können wir feststellen:

Selbst bei der Bewertung relativ einfacher und engumgrenzter Bewegungsabläufe ist die Objektivität und Reliabilität so niedrig, daß sie eher einen Rückschluß auf den Lehrer als auf die Leistung des Schülers zuläßt. Als Grundlage für eine Zensur sind deshalb solche „qualitativen Bewertungen" (im Sinne WIEGANDS, s. o.) völlig untauglich.

Ich glaube, wir dürfen dieses Ergebnis auf alle Sportarten, in denen nicht eindeutig gemessen werden kann, ausdehnen.

5. Messen im Verhaltensbereich

Sowohl Theoretiker als auch Praktiker sind sich weitgehend darüber einig, daß es weder dem Sinn nach der Zielsetzung des Schulsports entspräche, wenn die meßbare motorische Leistung alleinige Grundlage der Zeugniszensur wäre.

So wird auch in den Rahmenrichtlinien zum Sportunterricht der meisten Bundesländer ausdrücklich festgestellt, daß bei der Zensurenfindung die Veranlagung des Schülers, sein Leistungswille sowie sein soziales Verhalten mit zu berücksichtigen seien.

Um einige Beispiele zu nennen:

Am 3. November 1966 legte die KMK folgenden Beschluß vor:

„Die Wertungstabellen für die meßbaren Leistungen dürfen deshalb nicht der einzige Maßstab für die Beurteilung sein. Die Leistungsbereischaft, der Leistungswille und das Verhalten des Schülers in der übenden Gemeinschaft sind auch in ihrer Abhängigkeit von der konstitutionellen Veranlagung und dem Gesundheitszustand zu berücksichtigen."

Im „Lehrplan Sport für die Schulen Nordrhein-Westfalens" (o. J., ca. 1973) für die Grundschule wird gefordert:

„Die Zeugnisnote wird sowohl auf Leistungsmessungen in Form von Wettkämpfen und Tests als auch auf Beobachtungen von taktischem und kooperativem Verhalten sowie von Einsatz und Lernbereitschaft beruhen. ... Eine solche Bewertung bezieht die individuelle körperliche Lernvoraussetzung des einzelnen Schülers mit ein. Jede Beurteilung sollte auch motivierend und fördernd im Sinne des Strebens nach neuen Lernzielen sein" (Seite 10).

Für die Sekundarstufe I heißt es: „Die folgenden Empfehlungen müssen als erster Versuch angesehen werden, die Notengebung im Sport mehr zu vereinheitlichen. ... Die gemessenen Leistungen bilden nicht den einzigen Maßstab der Beurteilung. Folgende Komponenten sind angemessen einzubeziehen:

a) der individuelle Leistungsfortschritt
b) soziales Handeln (Vorbereitung von Wettkämpfen, Kampfrichtertätigkeit, Mitgestaltung des Unterrichts, Helfen und Sichern, Arbeit in der Gruppe, Teilnahme an schulischen Wettkämpfen u. a.)" (Seite 23).

Hamburg fordert 1973 in den „Richtlinien und Lehrplänen für die Grundschule" die Berücksichtigung

a) „des Bewegungsvermögens,
b) psychomotorischer Fähigkeiten,
c) der physischen Entwicklung,
d) schöpferischer Fähigkeiten,
e) des affektiven und sozialen Verhaltens (z. B. anpassungsfähig, selbständig, einsatzfreudig, ängstlich, hilfsbereit usw.)"