pedocs
Open Access Erziehungswissenschaften

DIPF

Borich, Gary D.; Klinzing, Hans G.

Die Beobachtung von Unterrichtsprozessen. Einige Vorannahmen und Empfehlungen zur Erfassung von Verhalten geringen Inferenzgrades

Unterrichtswissenschaft 14 (1986) 1, S. 65-79

Quellenangabe/ Reference:
Borich, Gary D.; Klinzing, Hans G.: Die Beobachtung von Unterrichtsprozessen. Einige Vorannahmen und Empfehlungen zur Erfassung von Verhalten geringen Inferenzgrades - In: Unterrichtswissenschaft 14 (1986) 1, S. 65-79 - URN: urn:nbn:de:0111-pedocs-295924 - DOI: 10.25656/01:29592

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-295924
https://doi.org/10.25656/01:29592

in Kooperation mit / in cooperation with:

BELTZ JUVENTA

http://www.juventa.de

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.
This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Digitalisiert

Mitglied der Leibniz-Gemeinschaft

Allgemeiner Teil

Gary D. Borich, Hans Gerhard Klinzing[1]

Die Beobachtung von Unterrichtsprozessen

Einige Vorannahmen und Empfehlungen zur Erfassung von Verhalten geringen Inferenzgrades

In diesem Beitrag werden vier Vorannahmen für die Beobachtung von unterrichtlichem Verhalten diskutiert und Vorschläge zur Anlage von Untersuchungen zu Unterrichtsprozessen gemacht. Den Instrumenten, die praktisch in jeder Untersuchung zum Lehrverhalten verwendet werden, liegen die folgenden Annahmen zugrunde: 1) das Verhalten, das beobachtet wird, tritt in ausreichender Häufigkeit auf, 2) ist repräsentativ für das Lehrer- oder Schülerverhalten, 3) die Anzahl der Beobachtungsgelegenheiten sollte für jedes Verhalten, das beobachtet wird, gleich sein, und 4) die Dauer des Zeitintervalls ist angemessen zur reliablen Erfassung des Verhaltens und sollte für jede Variable von gleicher Dauer sein. Empfohlen wird, daß die optimale *Dauer* einer Beobachtungsgelegenheit, ihre Anzahl und die *Dauer* des Zeitintervalls korrespondiert mit der Art des Verhaltens, das beobachtet werden soll. Deshalb sollten sie so eng wie möglich auf das zu bobachtende Verhalten abgestimmt werden.

The Observation of Classroom Processes. Some Assumptions and Suggestions for Improving Low-Inference Measurement

This paper presents four assumptions to the observation of classroom behavior and suggests some of the ways these assumptions should influence the design of a classroom observation study. Virtually every study of teacher behavior has used some type of classroom observation instrument. These instruments have assumed that 1) the behavior coded occurs frequently enough to be observed during the observation occasion, 2) the behavior coded is representative of the teacher's or pupil's typical behavior, 3) the number of observation occasions should be the same for each behavior measured, and 4) the duration of the time sampling interval is adequate to reliably detect the variable of interest and that this interval should be the same for each variable. This paper suggests that the optimal *length* of an observation occasion, *number* of observation occasions, and *size* of the time sampling interval vary according to the type of behavior being observed and, therefore, should be matched as closely as possible to that behavior.

In jüngerer Zeit ist die Methodologie der Unterrichtsbeobachtung gegenüber früheren Untersuchungen entscheidend verbessert worden. Heute enthält fast jede bessere Lehr-Lern-Studie objektive Beobachtungen von geringer Inferenz mit gut dokumentierten und akzeptierbaren Maßen für Beobachterreliabilität und Reliabi-

lität über Zeit. Darüber hinaus sind die eher intuitiven Vorstellungen davon, was gutes Unterrichten sei, ersetzt worden durch eine quantitative Erfassung von Schülerleistung. Das „Prozeß-Produkt-Paradigma"[2] hat frühere anekdotische Beschreibungen von Unterrichtsprozessen abgelöst.
Diese Änderungen haben bemerkenswerte Verbesserungen für die Methodik der Unterrichtsbeobachtung erbracht. Dennoch bleibt im Umfeld der Verfahren geringer Inferenz mit ihren an sich vielfältigen Vorteilen noch vieles offen, was zu erforschen, zu verstehen und zu verbessern ist. Dieser Beitrag soll mithelfen, den Blick zu schärfen für noch ungelöste methodische Probleme und für notwendige Verbesserungen. In diesem Sinne stellt er vier Vorannahmen dar, die in der Regel bei der Beobachtung von Unterrichtsprozessen getroffen werden und erörtert einige Möglichkeiten, wie diese bereits bei der Anlage einer Untersuchung berücksichtigt werden können.

Annahme 1: Der Beobachter nimmt an, daß das Verhalten, das er erfassen und kodieren will, in einem bestimmten Beobachtungszeitraum in ausreichender Häufigkeit auftritt

Die Intention, ein bestimmtes Verhalten zu beobachten, bedeutet noch nicht, daß dieses Verhalten im gegebenen Beobachtungszeitraum auch wirklich beobachtbar ist. In einer Reihe von Untersuchungen haben sich erhebliche Unterschiede in der Auftretenshäufigkeit bestimmten Verhaltens im Unterricht gezeigt, so daß die Möglichkeit besteht, daß einige der im Blick auf Schülerlernerfolg entscheidenden Verhaltensweisen nur selten vorkommen. So zeigte sich etwa in einer Untersuchung von *Moon* (1971), daß Lehrerfragen, die bei den Schülern das Erkennen von Zusammenhängen sowie das Aufstellen und Prüfen von Hypothesen erfordern, von Grundschullehrern sehr selten verwendet werden. Ähnliche Ergebnisse erbrachte eine Generalisierbarkeitsuntersuchung *(generazibility study)* zu einem weit verbreiteten Beobachtungsinstrument (*Erlich & Borich* 1979); hiernach traten 15% der erhobenen Variablen selten und inkonsistent auf oder waren bei weniger als 20% der Lehrer zu beobachten. Dasselbe zeigte sich bei zwei ganzen Gruppen von Variablen: „eigenständige Schülerfragen" und „eigenständige Schülerbeiträge".
Auf einen anderen Bereich, wo meist nur geringe Häufigkeiten der erhobenen Verhaltensweisen auftreten, verweist *Berliner* (1977), – den der Unterrichtsorganisation *(classroom management)*. Hier, so zeigt der Autor, sind manche Techniken so situations- oder schulspezifisch, daß sie in manchen Kontexten fast nie gebraucht werden (z.B. in Klassen mit Schülern von höherem sozioökonomischem Status mit überdurchschnittlicher Intelligenz), während in anderen Klassen ernste Organisationsprobleme ständig auftreten. Hinderlich bei der Untersuchung solcher Verhaltensweisen ist, daß es Lehrer geben mag, die fähig sind, Probleme so zu antizipieren bzw. diesen so vorzubeugen, daß manches Verhalten der Schüler erst gar nicht auftritt. Damit werden wichtige Praktiken, die ihrerseits bestimmte Arten von Verhalten hervorbringen oder auslöschen können, von der Beobachtung ausgeschlossen.

Eine Möglichkeit, wie theoretisch bedeutende, aber selten auftretende Verhaltensweisen erfaßt werden können, ist die Erhöhung der Dauer und Häufigkeit der Beobachtungen. Es ist zu vermuten, daß beides in einigen Untersuchungen zu unterrichtlichem Verhalten zur Erfassung bestimmter Verhaltensweisen oder zur Erreichung stabiler Schätzungen anderer nicht ausreichend war. Untersuchungen dazu von *Erlich, Borich* und *Bomberg* (1978) und *Erlich* und *Borich* (1979) zeigten große Unterschiede in der Anzahl der Beobachtungsgelegenheiten, die zur Erlangung von stabilen Schätzungen verschiedener, mit demselben Instrument gemessener Verhaltensweisen notwendig sind. In einer dieser Untersuchungen in Grundschulklassen im Fach Lesen (*Erlich, Borich* und *Bomberg* 1978) variierte die Anzahl der Beobachtungsgelegenheiten, die notwendig für eine Stabilitätsschätzung von 0.70 waren, von nur *einer* Beobachtungsgelegenheit von 50 Minuten Dauer („korrekte Schülerantworten auf Lehrerfragen") bis hin zu zehn fünfzigminütigen Beobachtungsgelegenheiten („Schüler bittet um Feedback vom Lehrer"). Zumeist haben Untersuchungen zu unterrichtlichem Verhalten jedoch weniger als 10 Beobachtungsgelegenheiten, einige sogar nur zwei (*Campbell* 1972). Die durchschnittliche Anzahl von Beobachtungsgelegenheiten in 12 Untersuchungen, von denen *Shavelson* und *Atwood* (1977) berichten, war fünf, und diese Anzahl stand im allgemeinen in einem Zusammenhang mit niedrigen Stabilitäts-Koeffizienten.

Eine Reihe von Untersuchungen zur Prüfung der Stabilität von Unterrichtsverhalten kann dazu dienen, die Dauer und Anzahl der Beobachtungsgelegenheiten, die zu einer zuverlässigen Messung bestimmter Arten von Verhalten notwendig sind, zu bestimmen. So fassen *Shavelson* und *Atwood* (1977) Stabilitätskoeffizienten für über 200 einzelne Unterrichtsprozeßvariablen zusammen. Ähnliche Informationen lassen sich entnehmen aus den bereits zitierten Untersuchungen von *Erlich* und *Borich* (1979) und *Erlich, Borich* und *Bomberg* (1978), – wenn auch für eine geringere Anzahl von Variablen.[3] Solche Untersuchungen dienen als Grundlage, um von Anfang an für die Verhaltensweisen, die für eine Untersuchung von Interesse sind, die jeweils angemessene Anzahl von Beobachtungsgelegenheiten vorzusehen. Wenn nun aber keine Befunde zur Stabilität oder Generalisierbarkeit vorliegen, wie kann der Forscher wissen, wie lang eine Beobachtungsgelegenheit sein muß und wie viele davon notwendig sind, um stabile Schätzungen zu erhalten? Wann zum Beispiel sollte eine Beobachtungsgelegenheit einen ganzen Schul-Vormittag oder im Unterschied dazu etwa nur eine 50minütige Unterrichtsstunde dauern, und weiterhin, wann erscheinen 10 Beobachtungsgelegenheiten angemessen oder etwa 5? Bei der Beantwortung dieser Fragen sind mindestens zwei Faktoren zu berücksichtigen. Zum einen ist dies die Zeit zwischen Anfang und Schluß der Beobachtung, in der mit hoher Wahrscheinlichkeit alle Informationen enthalten sind, die für die Diskrimination des Verhaltens relevant sind. Das zweite sind Kontext und Verhaltenssignale, die zusammen mit dem zu beobachtenden Verhalten auftreten. Darauf wird im folgenden einzugehen sein.

Es gibt keinen größeren Trugschluß im Bereich der direkten Verhaltensbeobachtung als den, daß auf einem Beobachtungssystem geringer Inferenz alle Verhaltensweisen wirklich vergleichbare Inferenzgrade repräsentieren, oder auch nur,

daß alle wirklich von geringer Inferenz sind. Eine Variable geringer Inferenz ist im allgemeinen definiert als spezifisch und diskret; das bedeutet, daß sie unabhängig von anderem Verhalten ist und eine zuverlässige Kodierung nicht von Kontext oder anderen Verhaltenssignalen abhängt. Der Grad der Inferenz eines Verhaltens steht im Zusammenhang mit der Dauer, die zu dessen Beobachtung benötigt wird. Dies wird deutlicher bei der Betrachtung globaler Schätzskalen von hoher Inferenz, mit denen eine relativ geringe Anzahl von Verhaltenskategorien beobachtet wird. Da Verhaltensweisen hoher Inferenz weniger diskret (mehr interdependent) und weniger spezifisch (aus mehr Komponenten zusammengesetzt) sind als die geringer Inferenz, kann ihre Kodierung nicht so schnell erfolgen. Das heißt, daß Verhaltensweisen hoher Inferenz sich im allgemeinen im selben Beobachtungszeitraum nicht so schnell wiederholen oder eine geringere „Umsatzrate" haben als Verhaltensweisen geringer Inferenz.

Da die Beobachtung mit Instrumenten hoher Inferenz immer ein Entfalten von mehreren Verhaltensfacetten bedeutet, ist mehr Zeit notwendig, um das Verhalten „sich ereignen" zu sehen. Zum Beispiel kann erwartet werden, daß zur Beobachtung von „Lehrerenthusiasmus" eine zeitlich längere Beobachtungsgelegenheit als zur Beobachtung von „Lehrervortrag" benötigt wird, da im ersten Fall notwendigerweise eine ganze Anzahl verschiedener Facetten des Verhaltens beobachtet werden muß, um zu einer zuverlässigen Einschätzung zu kommen. Auf der anderen Seite ist Verhalten geringer Inferenz anfälliger für falsches Kodieren, da der Unterrichts- und Verhaltenskontext, in dem es auftritt, schneller als bei Verhaltensweisen hoher Inferenz wechseln kann. Dies trifft insbesondere für solche Verhaltensweisen geringer Inferenz zu, deren Bedeutung leicht durch den Unterrichts- und Verhaltenskontext, in dem sie auftreten, beeinflußt wird.

Eine zweite Erwägung bezieht sich auf Kontext und Verhaltenssignale, die mit dem zu kodierenden Verhalten einhergehen. Diese Signale oder Zeichen repräsentieren die Intention des Lehrers, das soziale Klima im Unterricht, expressives und nichtverbales Verhalten sowie die Muster und Abfolgen, in denen das Verhalten auftritt. Im allgemeinen erfordert das zuverlässige Erfassen von Variablen hoher Inferenz die Berücksichtigung von weit mehr derartigen Signalen als das bei Variablen geringer Inferenz der Fall ist. Es ist das ganzheitliche Muster solcher Signale, das ein Verhalten hoher Inferenz grundsätzlich stabiler macht, da die Wahrscheinlichkeit, daß alle oder die meisten dieser Signale sich per Zufall ändern, geringer ist als bei einem spezifischen Verhaltenssignal, das ein Verhalten geringer Inferenz definiert. Daß viele Verhaltensweisen geringer Inferenz ohne solche Kontext- und Verhaltenssignale definiert sind, kann sie unstabiler machen als Verhaltensweisen hoher Inferenz. Dies dürfte speziell dann der Fall sein, wenn die Bedeutung eines Verhaltens geringer Inferenz leicht von Kontextsignalen beeinflußbar ist und wenn Regeln oder illustrierende Beispiele dafür, wie das Verhalten in unterschiedlichen Kontexten auftritt, nicht bereitgestellt werden. Damit wird das Argument plausibler, daß zur Erreichung etwa derselben Reliabilität für einige Kategorien geringer Inferenz mehr Beobachtungsgelegenheiten notwendig werden als für Einschätzungen hoher Inferenz (*Shavelson & Atwood* 1977). Und dies, obwohl diese Verhaltens-

weisen dazu tendieren, höhere „Umsatzraten" zu haben und folglich in einer gegebenen Zeitspanne häufiger beobachtet werden können.

Wie können diese beiden Überlegungen aber in eine Strategie eingehen, mit der die angemessene Dauer und Anzahl von Beobachtungsgelegenheiten zu bestimmen sind? Es zeigt sich, daß 1) für Verhalten geringer Inferenz eine kürzere Dauer der Beobachtung ausreichend ist als für Verhalten hoher Inferenz, da die „Umsatzrate" höher ist, und daß 2) für Verhalten geringer Inferenz eine größere Anzahl von Beobachtungsgelegenheiten erforderlich sein kann als für Variablen höherer Inferenz, da Kontext- und Verhaltenssignale, die ein Verhalten stabilisieren können, allgemein nicht in die Definition einer Variable geringer Inferenz eingehen. Eine Ausnahme mag sein, wenn Grundregeln zur Beobachtung oder Anhaltspunkte zur Beobachtung des Verhaltens als Hilfe zur Klassifizierung in die Kodierungsmanuals bzw. in das Beobachtertraining aufgenommen werden. Solche Grundregeln oder Anhaltspunkte sind etwa Lehreräußerungen und Schülerreaktionen, die helfen, ein bestimmtes Verhalten von anderen zu unterscheiden. Je mehr solcher Regeln und Hilfen zur Beobachtung eines Verhaltens aufgestellt sind, desto geringer kann die Anzahl der benötigten Beobachtungsgelegenheiten sein.[4] Je nachdem, ob das Verhalten von höherer oder geringerer Inferenz ist und Grundregeln oder andere Signale, die bei der Klassifizierung des Verhaltens helfen sollen, spezifiziert sind, kann das zu beobachtende Verhalten in vier Kategorien eingestellt werden, wie Tafel 1 zeigt:

Tafel 1: Empfehlungen für Dauer und Anzahl der Beobachtungsgelegenheiten für vier Kategorien von Verhalten

	Grundregeln oder andere Signale, die zur Klassifizierung eines Verhaltens nützlich sind, sind spezifiziert (Variablendefinition ist umfassend)	Grundregeln oder andere Signale, die zur Klassifizierung eines Verhaltens nützlich sind, sind nicht spezifiziert (Variablendefinition ist knapp)
Verhalten geringerer Inferenz (hohe „Umsatzrate"), meistens unabhängig von anderem Verhalten; erfordert vom Beobachter nur minimale Informationsverarbeitung oder Abstraktion	Relativ kurze Beobachtungsgelegenheiten und relativ wenige Beobachtungsgelegenheiten erforderlich* A	Relativ kurze Beobachtungsgelegenheiten und relativ viele Beobachtungsgelegenheiten erforderlich* B
Verhalten höherer Inferenz (geringe „Umsatzrate"), nicht unabhängig von anderem Verhalten; erfordert vom Beobachter beträchtliche Informationsverarbeitung und Abstraktion	Relativ lange Beobachtungsgelegenheiten und relativ wenige Beobachtungsgelegenheiten erforderlich* C	Relativ lange Beobachtungsgelegenheiten und relativ viele Beobachtungsgelegenheiten erforderlich* D

* Es geht hier nicht darum, daß viele (oder wenige) Beobachtungsgelegenheiten notwendig sind, sondern daß dies *relativ* zu anderen Feldern der Tafel der Fall ist.

Wenn zum Beispiel die Dauer eines Beobachtungszeitraumes auf eine 50minütige Unterrichtsstunde und die Anzahl der Beobachtungen auf fünf festgelegt wurden, können sich die Stabilitätskoeffizienten für die verschiedenen Kategorien des Verhaltens in den vier Feldern der Tafel 1 folgendermaßen unterscheiden: Die höchsten Stabilitätskoeffizienten können für Verhalten geringer Inferenz erwartet werden, zu dem Grundregeln der Beobachtung oder andere Verhaltenssignale zur Klassifizierung spezifiziert sind (Feld 1 in Tafel 1). Die nächsthöheren Stabilitätskoeffizienten sind in Feld 2 und 3 zu erwarten, wo zwar die optimale Dauer der Beobachtungsgelegenheiten erreicht sein kann, wo jedoch ihre Anzahl zu gering ist (Feld 2), bzw. wo ihre optimale Anzahl erreicht sein kann, ihre Dauer jedoch zu gering ist (Feld 3), um etwa denselben Grad von Stabilität wie in Feld 1 zu erlangen. Der geringste Stabilitätskoeffizient wird in Feld 4 zu erreichen sein, wo Anzahl und Dauer der Beobachtungsgelegenheiten zu gering sind.

Ziel einer solchen Kategorisierung des Verhaltens ist, die Dauer und Anzahl der Beobachtungsgelegenheiten in jedem Feld so abzustimmen, daß ein zufriedenstellender Grad von Stabilität erreicht wird. Die Beobachter können dann die zu jedem Feld der Tafel gehörenden Verhaltensweisen separat beobachten. Zugleich können einige Verhaltensweisen möglicherweise von Feld 4 nach oben in Feld 2 und 3 und von Feld 2 und 3 in Feld 1 eingereiht werden, indem die Definitionen des Verhaltens revidiert und/oder die Grundregeln bzw. Signale, die für die Klassifizierung des Verhaltens nützlich sind, neu formuliert werden, um eine höhere Stabilität bei geringerer Anzahl oder Dauer der Beobachtungsgelegenheiten zu erhalten.
Einige Beispiele aus der Untersuchung von *Borich* et al. (1977), die in die obere und untere Hälfte von Tafel 1 fallen, zeigt Tafel 2:

Tafel 2: Beispiele für Verhaltensweisen höherer und geringerer Inferenz in Beobachtungsinstrumenten „geringer Inferenz" mit dazugehörigen Generalisierbarkeits-Koeffizienten *(g)* aus: *Borich* et al. 1977)

Geringere Inferenz	*g*	Höhere Inferenz	*g*
Keine Kommunikation (CCO)	NV (1)	Gibt Analyse (CCO)	0.70 (5)
L. Gibt Sachinformationen zum spezifischen Unterrichtsgegenstand (CVC)	NV	Fragt nach Begründung, Erklärung, Einschätzung, Interpretation oder Beurteilung (CVC)	0.31 (2)
Lehrervortrag (FIAS)	0.86 (2)	Akzeptiert begründete Gedanken, Interpretationen, Einschätzungen (CVC)	0.35 (2)
L. Gibt einfache Sachinformationen (STARS)	0.72 (2)	Akzeptiert oder verwendet Schülergedanken (FIAS)	0.17 (2)
L. Erteilt Anweisungen (FIAS)	NV	Akzeptiert Gefühle (FIAS)	0.64 (2)
Keine Reaktion (B-G)	0.70 (3)	Auf Unterrichtsgegenstand bezogene Fragen, die auf den Erfahrungshintergrund des Schülers zielen (B-G)	0.70 (2)
Stellt neue Frage (B-G)	0.70 (4)		
Warnungen des Lehrers (B-G)	0.70 (4)	Überreaktion des Lehrers (B-G)	0.70 (6)
Lob auf richtige Antworten (B-G)	0.70 (2)		

CCO: Classroom Communication Observational System (Withall, Newell and Lewis, 1961)

1) Zu geringe oder negative Varianz, um zuverlässige Schätzung zuzulassen

CVC: CERLI Verbal Behavior Classification System (Cooperative Education Laboratory, 1969)

FIAS: Flanders Interaction Analysis System (Flanders, 1970)

STARS: Spaulding Teacher Activity Rating Schedule (Spaulding, 1967)

B-G: Brophy-Good Dyadic Interaction System (Brophy and Evertson, 1976)

2) 3 Beobachtungsgelegenheiten und 2 Beobachter
3) 9 Beobachtungsgelegenheiten und 2 Beobachter
4) 4 Beobachtungsgelegenheiten und 2 Beobachter
5) Kombination von 8 oder weniger Beobachtern und 8 oder weniger Beobachtungsgelegenheiten
6) 7 Beobachtungsgelegenheiten und 2 Beobachter

In Tafel 2 zeigt sich, daß die Stabilitäts- oder Generalisierbarkeitskoeffizienten für Verhalten stark variieren, obwohl sie aus „Beobachtungsinstrumenten geringer Inferenz" gewonnen wurden. Dies läßt die Vermutung zu, daß eine 50minütige Beobachtungsgelegenheit angemessen ist, wenn Verhalten geringerer Inferenz kodiert werden soll und daß bei Verhalten höherer Inferenz eine Beobachtungsdauer von einem halben Tag angezeigt ist. Ähnlich darf vermutet werden, daß bei Variablendefinitionen, die zusätzlich Grundregeln und Aktions- und Reaktionssignale als Klassifikationshilfen bereitstellen, fünf Beobachtungsgelegenheiten angemessen sind, daß jedoch bei wenig umfassenden Variablendefinitionen (z.B. solchen, die im Kodierungsmanual nicht mit illustrierenden Beispielen oder ausreichenden Interaktionssequenzen versehen sind), zehn Beobachtungsgelegenheiten notwendig werden können. Obwohl dies nur Vermutungen sind, können sie als Ausgangspunkt für Generalizibility-Untersuchungen und vielleicht als Planungsrichtlinien für eine Beobachtungsuntersuchung dienen.

Annahme 2: Der Beobachter nimmt an, daß das beobachtete Verhalten repräsentativ für das Verhalten der Lehrer oder Schüler ist

Die Stabilität von unterrichtlichem Verhalten – oder die Zuverlässigkeit über Beobachtungsgelegenheiten hinweg – hängt von vielen Faktoren ab. In Tafel 3 werden solche Faktoren (Facetten) aufgeführt:

Tafel 3: Facetten, von denen ein Einfluß auf die Stabilität von Verhalten bekannt ist oder vermutet werden kann

Tageszeit (z.B. Vormittag vs. Nachmittag)
Unterrichtsfach (z.B. Englisch vs. Mathematik)
Spezifischer Unterrichtsinhalt (z.B. Wortschatz vs. Rechtschreibung)
Unterrichtsstil (z.B. direkte vs. indirekte Unterweisung)
Arrangement in der Klasse (z.B. festgelegte vs. flexible Sitzanordnung)
Ziele des Lehrers (z.B. Wiederholung vs. neuer Inhalt)
Gruppengröße (z.B. ganze Klasse vs. Kleingruppen)
Unterrichtsmethode (z.B. Lehrervortrag vs. Unterrichtsdiskussion)
Zeitpunkt im Schuljahr (z.B. zeitliche Nähe zu einem Feiertag oder den Ferien)
Zeitpunkt in der Woche (z.B. Freitag vs. Montag)
Schulstufe (z.B. Grundschule vs. Mittelstufe
Schule (z.B. Innercity vs. Suburban)
Stundenpläne für Parallelklassen oder für aufeinander folgende Jahre (z.B. Unterricht im selben Fach am gleichen Tag oder in aufeinanderfolgenden Schuljahren oder -halbjahren)

Von diesen kontextbezogenen Variablen kann erwartet werden, daß sie die Stabilität der zu beobachtenden Verhaltensweisen beeinflussen. Diese Variablen werden auch „Facetten" genannt, weil das innerhalb jeder Facette auftretende Verhalten jeweils nur *eine* Schätzung des Verhaltens repräsentiert, wenn es in einem größeren Kontext beobachtet wird.

Wenn für unterrichtliches Verhalten bekannt ist oder doch vermutet werden kann, daß es über die Elemente einer Facette hinweg variiert, müssen die Beobachtungen über eine hinreichende Anzahl von Beobachtungsgelegenheiten hinweg durchgeführt werden, damit sie alle Elemente einer Facette repräsentieren, – oder aber es kann eine Verallgemeinerung von Befunden nur für die Elemente einer Facette erfolgen, die wirklich beobachtet worden sind (z.B. Unterricht am Vormittag, nicht jedoch am Nachmittag). Im ersten Fall kann dies so viele Beobachtungsgelegenheiten erfordern, daß es leicht den Aufwand, der für die Unterrichtsbeobachtung möglich ist, übersteigt. Bei vielen Facetten kann jedoch so verfahren werden, daß manche Elemente einfach unberücksichtigt gelassen oder als unwichtig ausgeschlossen werden. Zum Beispiel sind Schulen und Klassen oft von vornherein festgelegt, wenn sie als „repräsentativ", als „Follow Through"[5] oder als Schulen mit Schülern von Eltern mit geringem sozioökonomischem Status für eine Untersuchung ausgewählt worden sind. Ähnlich sind der Inhalt bestimmter Unterrichtsfächer und die Klassenstufen dadurch festgelegt, daß einfach die Ergebnisse der Untersuchung nur auf bestimmte Unterrichtsfächer oder -gegenstände bzw. Klassenstufen bezogen werden. Einige Facetten, die zunächst bedeutsam erscheinen, erweisen sich als unwichtig, wenn andere Facetten festgelegt werden. Zum Beispiel wird angenommen, daß beträchtliche Unterschiede im Schülerverhalten zwischen Vormittags- und Nachmittagsunterricht bestehen. Aber die Unterrichtsfächer Lesen und Rechnen sind in den ersten Klassen (in den USA) fast immer auf den Vormittag beschränkt. Somit sind möglicherweise der Zeitpunkt innerhalb der Woche und der innerhalb des Schuljahres die einzigen Variablen, für die es nötig ist, die Beobachtungsgelegenheiten nach Zufall anzuordnen. Überdies können auch innerhalb dieser Facetten einige Zeitpunkte als weniger repräsentativ für ein typisches Auftreten von bestimmten Verhaltensweisen angesehen werden als andere. So kann etwa der Freitagnachmittag (wenn sonnabends keine Schule ist) ebenso wie auch der Anfang oder das Ende eines Schuljahres und die Zeit vor Beginn oder nach Ende der Ferien ausgelassen und die Unterrichtszeit, auf die die erforderlichen Beobachtungsgelegenheiten verteilt werden, eingeengt werden.

Solche Facetten können eine bedeutsame Rolle für die Stabilität eines Verhaltens spielen.

In der *Texas Teacher Effectiveness Study* etwa (*Brophy, Evertson, Crawford, King* und *Senior* 1975; *Crawford, Brophy* und *Evertson* 1975) waren die Stabilitätskoeffizienten innerhalb eines Jahres und über zwei aufeinanderfolgende Jahre hinweg für Lehrer in der zweiten und dritten Klasse recht niedrig, von etwa 0.10 bis 0.30. In einer Nachfolgeuntersuchung, in der Lehrer in der 7. und 8. Klasse (Junior High School) in Parallelklassen über ein einziges Jahr unterrichteten, waren die Stabilitätskoeffizienten für viele derselben Variablen recht hoch, von etwa 0.70 bis 0.90 (*Evertson, Anderson* und *Ecker* 1977). In der ersten Untersuchung waren die Stabilitätskoeffizienten über drei unterschiedliche Facetten niedrig: Vormittagsunterricht vs. Nachmittagsunterricht, Unterricht in der ganzen Klasse (vormittags) vs.

Unterricht in Lese-Kleingruppen und Unterricht in ganzen Klassen (nachmittags) vs. Lese-Kleingruppen. Obwohl die Stabilitätskoeffizienten im zweiten Jahr der Untersuchung leicht höher lagen (hier war die Anzahl der Beobachtungsgelegenheiten höher), wiesen die Ergebnisse insgesamt in dieselbe Richtung. Damit zeigte sich der Einfluß dieser Facetten auf die Stabilität des in dieser Untersuchung erhobenen Verhaltens.
Im Unterschied zu dieser Untersuchung wurde in der *Texas Junior High School Study* die Stabilität von vielen dieser selben Variablen für Lehrer, die in Parallelklassen in Englisch und Mathematik unterrichteten, geprüft. Während die Stabilitätskoeffizienten *über* diese Unterrichtsfächer hinweg nicht sehr hoch waren, ergaben sich dagegen sehr hohe Werte für die Stabilität *innerhalb* der Unterrichtsfächer über Klassen hinweg. Damit zeigte sich der bedeutsame Einfluß des Unterrichtsfaches auf die Stabilität des in der Untersuchung erhobenen Verhaltens.
Da außerdem die Lehrer am selben Tag in Parallelklassen unterrichteten (im Unterschied zu der früheren Untersuchung, wo vergleichbare Klassen in zwei aufeinanderfolgenden Jahren unterrichtet wurden), können einige der Unterschiede in der Stabilität des Verhaltens in den beiden Untersuchungen gut eine Funktion von zeitlichen Differenzen (parallele versus nacheinander erfolgte Erhebung der Daten) sowie von durch sie bedingten Differenzen in der inhaltlichen Planung und Organisation sein.

Worauf es hier ankommt, ist zu zeigen, daß Variablen geringer Inferenz besonders dann in Verdacht stehen, instabil zu sein, wenn Facetten, wie sie in Tafel 2 aufgelistet sind, unbeachtet bleiben. Denn diese repräsentieren eine Vielfalt von Zeichen und Signalen aus dem Unterrichtskontext heraus, die die Bedeutung eines Verhaltens beeinflussen können.

Annahme 3: Der Beobachter nimmt an, daß die Anzahl der Beobachtungsgelegenheiten für jedes einzelne Verhalten gleich zu sein hat

Leider sind nur wenige Untersuchungen bekannt, in denen die Anzahl der Beobachtungsgelegenheiten auf die Art des beobachteten Verhaltens abgestimmt wurde, wie dies in Tafel 1 empfohlen wird. Im allgemeinen wird vor Beginn einer Untersuchung eine bestimmte Anzahl von Beobachtungsgelegenheiten bestimmt und angenommen, daß sie für alle in Frage stehenden Verhaltensweisen angemessen ist. Die für die Stabilität eines Verhaltens notwendige Anzahl der Beobachtungsgelegenheiten kann jedoch, wie dies aus den Generalisierbarkeits-Koeffizienten *(generalizibility coefficient)* in Tafel 2 hervorgeht, beträchtlich variieren. Wenn nicht häufig genug beobachtet wird, kann es vorkommen, daß ein Verhalten überhaupt nicht erfaßt wird; wird zu häufig beobachtet, so wird wertvolle Beobachtungszeit verschwendet. Ziel einer Generalisierbarkeits-Studie ist es zu bestimmen, wieviele Beobachtungsgelegenheiten und Beobachter nötig sind, um stabile Schätzungen eines Verhaltens zu erhalten. Einmal abgesehen davon, daß nur eine ungenügende Anzahl solcher Untersuchungen vorliegt, so sind die Ergebnisse der vorhandenen Untersuchungen viel zu wenig berücksichtigt worden.
Eine Möglichkeit, wie die Ergebnisse solcher Untersuchungen Verwendung finden können, ist die Festsetzung einer Mindestanzahl von Beobachtungsgelegenheiten und Beobachtern, die für die Erreichung einer Stabilität von etwa 0.70 für alle zu beobachtenden Verhaltensweisen einer Untersuchung notwendig ist. Dies setzt natürlich voraus, daß für jede Verhaltensweise einer Untersuchung Generalisierbarkeits-Studien vorliegen, die Generalisierbarkeits-Koeffizienten *(g)* liefern; dies aber ist selten der Fall. Auch wenn *g*-Koeffizienten errechnet sind, dienen sie selten

als Grundlage für die Festlegung der Anzahl der Beobachtungsgelegenheiten. Wie viele Beobachtungsgelegenheiten festgelegt werden, scheint bis heute eine Sache der Intuition und Vermutung zu sein, genauso wie zu der Zeit, bevor die einschlägigen Publikationen von *Cronbach, Glaser, Nada* und *Ratjaratnam* (1972), *Shavelson* und *Atwood* (1977) sowie *Borich, Calkins, Malitz, Erlich, Kugle* und *Pascone* (1977) erschienen waren, die Methodologie und Befunde zu diesem Thema liefern.

Schließlich sei hier angemerkt, daß, selbst wenn um der praktischen Durchführung einer Untersuchung willen eine gleichbleibende Anzahl von Beobachtungsgelegenheiten für alle zu beobachtenden Verhaltensweisen richtig erscheinen mag, man sich doch nicht notwendigerweise einer solchen Logik beugen muß. Wie gut wir wirklich beobachten, ist relativ wenig erforscht, hat jedoch wichtige Implikationen für die Stabilität der erfaßten Verhaltensweisen. Bei der Einschätzung der Kosten für eine Beobachtung, in der die Verhaltensweisen öfter beobachtet werden als möglicherweise notwendig, wird man zum Beispiel finden können, daß, wenn die Beobachtung eines bestimmten Verhaltens beendet wird, da eine ausreichende Stabilität erlangt ist, die Genauigkeit, mit der die verbleibenden Verhaltensweisen beobachtet werden, sogar steigt. Hinzu kommt, daß es keinen Grund gibt, warum unterschiedliche Verhaltensweisen nicht in einer unterschiedlichen Anzahl von Beobachtungsgelegenheiten von unterschiedlichen Beobachterpaaren beobachtet werden könnten. Dies kann zu besseren Stabilitätsschätzungen führen als wenn – wie dies aus einigen Generalisierbarkeits-Untersuchungen hervorgeht – die Anzahl der Beobachter erhöht wird. Mit anderen Worten, anstatt mit mehr Beobachtern dasselbe Verhalten beobachten zu lassen, kann es, um angemessene Stabilitätsmaße zu erhalten, vernünftiger sein, die Beobachterpaare auf unterschiedliche Verhaltensweisen zu verteilen, die – wie in Tafel 1 empfohlen – eine unterschiedliche Anzahl von Beobachtungsgelegenheiten erforderlich machen.

Annahme 4: Der Beobachter nimmt an, daß die gewählte zeitliche Erstreckung des Beobachtungsintervalls zur zuverlässigen Erfassung einer Variablen angemessen ist und daß sie für jede Variable gleich sein sollte

Diese Annahme zielt auf Unterschiede in der Dauer eines Beobachtungsintervalls. Ein Beobachtungsintervall ist definiert als die Zeitspanne, die für eine einzelne abgegrenzte Beobachtung und ihre Kodierung vorgesehen ist. Die Verfahren zur Aufzeichnung des Beobachteten können sich jedoch im Blick auf die Dauer eines solchen Intervalls stark unterscheiden. Die Verwendung von Merkmalssystemen,[6] die Protokollierung des Auftretens eines Verhaltens innerhalb eines bestimmten Zeitabschnittes, kann, obwohl sehr häufig angewandt, abhängig von der Dauer des zeitlichen Beobachtungsabschnitts äußerst unterschiedliche Ergebnisse bringen. Für einige Merkmalssysteme werden relativ lange (etwa 15 Minuten oder länger,)[7] für andere wiederum relativ kurze Beobachtungsintervalle (z.B. 3 Sekunden,)[8] verwendet. Je weniger diskret und spezifisch ein Verhalten ist (bei Verhalten höheren Inferenzgrades also), desto länger sollte das Zeitintervall sein. Umgekehrt sollte, je mehr diskret und spezifisch ein Verhalten ist (ein Verhalten geringeren Inferenzgrades also), das Zeitintervall entsprechend kürzer sein. Dieser Hinweis folgt der Erkenntnis, daß, wenn ein Verhalten häufig und unabhängig von anderem Verhalten auftritt, ein kurzes Beobachtungsintervall ausreicht, um einen häufigen Wechsel seiner Erscheinungsform einzufangen. Wenn ein Verhalten in relativ geringer Häufigkeit auftritt und/oder seine Erfassung von anderen Ereignissen beeinflußt wird (wie etwa die Intention des Lehrers, expressives nichtverbales Verhalten oder Verhaltensmuster bzw. -abfolgen), so wird ein längeres Zeitintervall notwendig, um seine An- oder Abwesenheit zu erfassen.

Ein wichtiger, jedoch selten diskutierter Aspekt im Zusammenhang mit der zeitlichen Erstreckung eines Beoachtungsintervalls ist der Grad, in dem der Beobachter innerhalb eines Beobachtungsintervalls Informationen abstrahiert und integriert, oder aber das Verhalten unmittelbar in dem Augenblick, in dem er es beobachtet, aufzeichnet. Bei Beobachtungssystemen mit lang dauernden Intervallen (in der Größenordnung von 10 Minuten oder länger) wird allgemein versucht, dasjenige Verhalten zu kodieren, das das gesamte Intervall am *besten* beschreibt. Andererseits wird bei Instrumenten mit kurzen Beobachtungsintervallen (in der Größenordnung von einer Minute oder weniger) – ob dies nun intendiert ist oder nicht – eher eine Technik der Momentaufnahme benützt, bei der keine oder wenig Informationsverarbeitung und -interpretation geschieht, außer der, die notwendig ist, um das Verhalten (unmittelbar) zu erkennen. Diese beiden Typen von Beobachtungsintervallen können völlig andere Bilder einer Interaktionssituation bieten, da im ersten Fall der Genauigkeitsgrad der Daten zumindest z.T. von der Fähigkeit des Beobachters abhängt, die Informationen während des Beobachtungsintervalls zu verarbeiten und zu integrieren, im zweiten Fall von der Fähigkeit des Beobachters, das Verhalten quasi automatisch und unmittelbar zu erkennen. Obwohl beide Typen bei statistischer Verarbeitung einer genügenden Anzahl von Beobachtungsintervallen ein repräsentatives Bild der unterrichtlichen Interaktionen liefern können, birgt bei längeren Beobachtungsausschnitten die kognitive Integration des Beobachteten, die der statistischen Verarbeitung vorausgeht, reichlich Gelegenheit zu Fehlern, wenn die Wahrnehmungen der Kodierer anfällig sind für Verzerrungen.

Zu einer Reihe von Pilotuntersuchungen z.B., die von dem ersten Autor durchgeführt wurden, wurden zwei nach Zufall zusammengesetzte Gruppen von Beobachtern in die Verwendung der Interaktionsanalyse nach *Flanders* (FIAS) (*Flanders* 1970) eingeübt. Der ersten Gruppe wurde auf Videotape die Unterrichtsstunde eines Lehrers gezeigt, in der Tische und Stühle in der Klasse absichtlich unordentlich aufgestellt wurden, die Anschlagbretter schief aufgehängt wurden und Papier auf den Boden verstreut wurde, um den Eindruck zu erwecken, daß die Klasse außer Kontrolle sei. Sie war es jedoch nicht, denn der Unterricht fand nach einem sorgfältig ausgearbeiteten Drehbuch statt, in dem die Schüler gemäß vorheriger Absprache gut mitspielten. Diese Unterrichtsstunde wurde exakt mit denselben Schülern für eine zweite Gruppe von Kodierern wiederholt, die bis zum selben Standard von Zuverlässigkeit trainiert worden war wie die erste Gruppe. Doch dieses Mal waren Tische und Stühle ordentlich in Reihen aufgestellt, die Anschlagbretter geradè ausgerichtet und abwechslungsreich gestaltet und der Boden war von Papier und Abfall gesäubert. Der Unterricht wurde exakt genauso durchgeführt wie beim ersten Mal und die schauspielenden Schüler reagierten genauso wie zuvor. Es ergaben sich jedoch zwischen den beiden Gruppen von Kodierern erhebliche Unterschiede zwischen den Kodierungen bei mehreren Kategorien des Flanders-Systems, obwohl dieses lediglich die *unterrichtliche Interaktion* erfaßt. Signifikante Unterschiede ergaben sich zugunsten des Unterrichts in der ordentlich geführten Klasse für die Häufigkeit von Lob (Kategorie 2) und dem Eingehen auf, bzw. Akzeptieren und Aufgreifen von Schülergedanken und -gefühlen (Kategorie 1 und 3 des FIAS). Die Kodierer waren offensichtlich von Dingen beeinflußt, die für das, was sie kodierten, irrelevant waren. Dies zeigt, daß Kodierer in dem, was sie wahrnehmen oder nicht wahrnehmen, vom Kontext, in dem der Unterricht stattfindet, beeinflußt werden können. Mit Recht kann vermutet werden, daß die Möglichkeit einer durch Kontext beeinflußten Wahrnehmungsverzerrung bei Merkmalen oder Kategorien, bei denen der Kodierer Informationen verarbeiten und integrieren muß, größer ist als bei denen, wo er das Verhalten unmittelbar aufzeichnet.

Unglücklicherweise ist bei einigen Beobachtungssystemen nicht angegeben, ob der Beobachter ein *gesamtes,* repräsentatives Bild des Beobachtungsintervalls oder eine Momentaufnahme erstellen soll. Eine gewisse Abstraktion und Interpretation

von Information wird immer erforderlich sein, wenn ein gesamtes, repräsentatives Bild eines Beobachtungsintervalls gewünscht wird, – dies ganz einfach deshalb, weil es für den Menschen unmöglich ist, das, was während eines Intervalls geschieht, vollständig wahrzunehmen und aufzuzeichnen, wie kurz es auch sein mag. Solche Abstrahierungen erfordern Urteile und diese wiederum beinhalten Wertungen. Wenn die dem Beobachtungsinstrument zugrundeliegenden und vom Beobachter zu treffenden Wertentscheidungen in der Beobachtungsanleitung oder während des Beobachtertrainings nicht thematisiert werden, werden unweigerlich die persönlichen Werte der Kodierer in die Kodierung mit eingehen.

Einige Beobachtungsinstrumente verwenden keine Zeitintervalle, sondern zeichnen jedes Verhalten oder jeden Verhaltenswechsel auf. Die Interaktionsanalyse von *Flanders* ist oft als 3-Sekunden-System mißverstanden worden.[9] Dabei soll in ihm der Wechsel im Verhalten erfaßt werden. Nur wenn ein Verhalten innerhalb einer Kategorie länger als drei Sekunden dauert, wird es alle 3 Sekunden kodiert. Wie auch eine Reihe anderer weit verbreiteter Beobachtungssysteme verzeichnet das Brophy-Good Dyadic Interaction System fortlaufend die Verhaltenswechsel. Damit sind bei diesen Systemen solche Probleme, wie sie oben angesprochen worden sind, vermieden.

Da im Augenblick viel diskutiert,[10] sollen hier noch einige der Grundannahmen zur Sprache kommen, die den Verfahren zur Auswahl der Zeitstichproben bei der Erfassung der Variablen: Zeit, in der die Schüler eine Aufgabe mit hoher Erfolgsrate lösen *(academic learning time),* aktive Lernzeit *(engaged time)* oder Aufmerksamkeitsrate *(rate of engagement)* zugrundeliegen. Die Beobachtungsinstrumente, die solche Variablen erfassen, messen nicht wirklich die Zeit oder wie lang ein solches Verhalten dauert. Bei ihnen werden vielmehr Intervallkodierungen in Form von Häufigkeiten in die tatsächliche Anzahl von Sekunden, Minuten oder Stunden umgewandelt, die diese Intervalle repräsentieren. Die Summe der Intervalle, in denen das Verhalten kodiert wurde, wird als Auftretenshäufigkeit der *Intervalle,* innerhalb derer das Verhalten beobachtet wurde, verzeichnet.[11]

Dieser Methode der Aufzeichnung und Verrechnung liegt die Annahme zugrunde, daß das jeweilige Verhalten charakteristisch ist für das gesamte Beobachtungsintervall. Wenn diese Annahme richtig ist, dann ist es möglich, Schlußfolgerungen in bezug auf die Zeit, den prozentualen Anteil an Zeit und andere Quotienten zur zeitlichen Erstreckung des beobachteten Verhaltens zu ziehen, immer im Vergleich zu der Zeit, in der das Verhalten hätte beobachtet werden können. Wenn solche Beobachtungsintervalle in zureichender Anzahl vorhanden und repräsentativ für Verhalten innerhalb eines Schultages und über Schultage hinweg sind, dann kann auf ein Schulhalbjahr oder sogar auf ein Schuljahr extrapoliert werden. Insofern als aufgrund der Erfassung von Verhalten durch ein Merkmalssystem Ergebnisse hinsichtlich der Zeit nicht aufgrund wirklicher zeitlicher Erstreckung eines Verhaltens, sondern aufgrund der Anzahl von Beobachtungsintervallen entstehen, in denen das Verhalten auftrat bzw. nicht auftrat, unterliegt es denselben Vorannahmen, wie sie für Merkmalssysteme gelten, die mit Häufigkeiten arbeiten. Das heißt, daß das beobachtete Verhalten innerhalb des Beobachtungsintervalls konstant bleibt und daß etwaige Irrtümer, die aus Veränderungen innerhalb des Intervalls erwachsen, ausgemittelt werden, vorausgesetzt, daß die Ergebnisse auf einer angemessen großen Anzahl von Intervallen beruhen. Mit anderen Worten: Es wird angenommen, daß

für jedes Intervall, für das ein Kodierer das Auftreten eines Verhaltens verzeichnet, das tatsächlich aber während eines größeren Teiles des Intervalls nicht auftrat, es ein vergleichbares Intervall gibt, für das der Kodierer das Nicht-Auftreten eines Verhaltens verzeichnet, das tatsächlich aber während eines größeren Teils des Intervalls vorhanden war. Damit geht man davon aus, daß Meßfehler zufällig sind. Wenn diese Annahme richtig ist, können kumulative Daten über Intervalle als *Schätzungen* der Zeit, die ein Verhalten andauert, genommen werden. Andernfalls sollte dies vermieden werden. Wenn Verhalten in nur wenigen Beobachtungsgelegenheiten beobachtet wurde oder bedeutsame Facetten nicht berücksichtigt wurden, dann ist es wahrscheinlich, daß diese Annahme *nicht* zutrifft.

Unterrichtswissenschaftler weisen gern auf technische und menschliche Probleme hin, die bei der Aufzeichnung der „aktiven Lernzeit" *(engaged learning time)* in ihrer *wirklichen* zeitlichen Erstreckung auftreten können. In Untersuchungen zum unterrichtlichen Verhalten kann diese jedoch eine bedeutende Rolle spielen; sie könnte darin bestehen, durch Pilotuntersuchungsdaten zu bestätigen, daß die Ergebnisse, die durch Aggregation von Intervallen einer bestimmten Länge gewonnen wurden, sich an die zeitliche Erstreckung des Verhaltens, die die Intervalle repräsentieren sollen, annähern. Dies würde bedeuten, daß für ausgewählte Unterrichtsausschnitte sowohl Methoden, die die wirkliche zeitliche Erstreckung eines Verhaltens erfassen, als auch Intervallmethoden verwendet werden und daß im Ergebnisvergleich die Größe der Diskrepanz zwischen beiden bestimmt wird. Große Diskrepanzen werden zur Vorsicht hinsichtlich der gewählten Dauer des Intervalls und/oder der zur Beobachtung ausgewählten Verhaltensweisen anhalten. In einem solchen Fall wird es ratsam sein, eine andere Intervalldauer zu wählen.

Schließlich ist die Frage zu stellen, ob für jedes beobachtete Verhalten das Beobachtungsintervall gleich lang sein soll. Dies ist praktisch bei jedem Beobachtungssystem der Fall, – wenn auch nur wegen des übermäßigen technischen Aufwandes, der bei Verwendung von unterschiedlichen Beobachtungsintervallen in ein und demselben Instrument entsteht. Obwohl es keine praktische Alternative zu diesem Vorgehen gibt, sollte doch beachtet werden, daß ein Beobachtungsinstrument geringen Inferenzgrades sowohl Variablen geringerer als auch höherer Inferenz enthalten kann. Da bei Variablen höherer Inferenz zumeist Kontextsignale und andere Anhaltspunkte zur Klassifizierung in ihrer operationalen Definition enthalten sind, bedarf es eines längeren Beobachtungsintervalls zur zuverlässigen Feststellung und Kodierung des Verhaltens als bei Variablen geringer Inferenz. Zum Beispiel erfordert ein Verhalten wie „Lehrerfrage" ein kürzeres Beobachtungsintervall als die Kodierung: „Schüler ist aufmerksam" (pupil is on-task): im ersten Beispiel kann das Verhalten allein aufgrund der Dauer eines einzigen vom Lehrer formulierten Satzes zuverlässig kodiert werden, während beim zweiten Beispiel die Beobachtung von vielerlei Signalen wie Gesichtsausdruck, Blickkontakt, Körperhaltung und das Ansehen der Arbeiten des Schülers bei der Stillarbeit weit länger dauern kann. Wenn die Dauer des Beobachtungsintervalls für den ersten Fall angemessen ist, kann sie nicht für den zweiten angemessen sein. Somit kann für das zweite Verhalten eine weniger stabile Schätzung erwartet werden, da die Kodierer ihre Beobachtung einfach auf das, was durch das Beobachtungsintervall bestimmt ist, begrenzen, – wenn sie nicht unbemerkt das Beobachtungsintervall ändern. Eine rechtzeitige Pilotuntersuchung zu jedem zu kodierenden Verhalten kann daher nützlich sein, bevor die Dauer eines Beobachtungsintervalls festgelegt wird; dieses

kann dann so angelegt werden, daß es der höchsten Dauer, die für die Wahrnehmung und Aufzeichnung der zu beobachtenden Verhaltensweisen erforderlich ist, angepaßt ist.

Schlußbemerkung

Die Verwendung von Beobachtungsinstrumenten geringer Inferenz bietet gegenüber subjektiven Schätzmethoden zahlreiche Vorteile, so daß die Gefahr besteht, daß Meßprobleme und Vorannahmen, die diesem Verfahren unterliegen, aus den Augen verloren werden. Dabei kann es leicht geschehen, daß man sich auf die Vorteile solcher Verfahren verläßt, ohne eine angemessene Behutsamkeit gegenüber ihren Schwächen walten zu lassen. Wir glauben jedoch, daß es nur aufgrund einer offenen Diskussion der Stärken und Schwächen solcher Beobachtungsmethoden eines Tages möglich sein wird, die angemessene *Dauer* einer Beobachtungsphase, die *Anzahl* der Beobachtungsgelegenheiten und die *Größe* eines Beobachtungsintervalls auf das zu messende Verhalten abzustimmen und praktikable Wege zu finden, diese Anforderungen in die Untersuchung von Unterrichtsprozessen einzubringen.

Anmerkungen

[1] Der Beitrag wurde während der Gastprofessur von *G. D. Borich* am Zentrum für Neue Lernverfahren, Universität Tübingen, verfaßt.

[2] Zum *Prozeß-Produkt-Paradigma* in der Unterrichtsforschung vgl. *Gage* 1963; 1978, *Doyle* 1978.

[3] Gemessen mit dem Brophy-Good Teacher-Child Dyadic Interaction System, dem Spaulding Teacher Activity Rating Schedule (STARS), dem Observation Schedule and Record (OSCAR 5) und dem Flanders Interaction Analysis System (FIAS).

[4] Als Beispiel für solche Grundregeln sei eine genannt, die Flanders seinem System (FIAS, *Flanders* 1970) beigegeben hat: Die Beobachter erhalten u.a. die Instruktion, die Kategorie 4 (Lehrerfrage) nicht zu verwenden in Fällen, wo der Lehrer vom Schüler keine Antwort erwartet; dies ist u.a. der Fall bei rhetorischen Fragen.

[5] *Follow-Through-Programme* sind Programme, die von der Bundesregierung der Vereinigten Staaten von Amerika seit 1967 gefördert werden und darauf zielen, speziellen Unterricht für Kinder aus Familien mit geringem Einkommen bereitzustellen, die an „Headstart" oder ähnlichen Vorschulprogrammen teilgenommen haben. Ziel des Gesamtprogramms war, verschiedene einzelne Programme aufgrund verschiedener Erziehungstheorien zu entwickeln und zu implementieren.

[6] Zum Unterschied von Merkmals- und Kategoriensystemen vgl. *Schulz* et al. (1970), Sp. 770ff.

[7] Z.B. das „Florida Climate Control System (FLACCS)" (*Soar* 1975) oder die Classroom Observation Scale (COS) (*Emmer* 1978).

[8] Z.B. das „Technion Diagnostic System" (*Bar-On & Perlberg* 1976; *Klinzing & Bar-On* 1975.

[9] Z.B. von *Hanke, Mandl & Prell* 1973; *Merkens & Seiler* 1978.

[10] S. vor allem: *Denham & Liebermann* (Hrsg.) 1980.

[11] Vgl. Beginning Teacher Evaluation Study: Observer's Manual. Oktober 1976.

Literatur

Bar-On, E. & Perlberg, A.: Der Facetten-Ansatz bei der Entwicklung einer Unterrichtstheorie. In: *Zifreund, W.* (Hrsg.): Training des Lehrverhaltens und Interaktionsanalyse. Weinheim & Basel 1976, 563–605.

Beginning Teacher Evaluation Study: Observer's Manual. O. O. October 1976.

Berliner, D.: Impediments to measuring teacher effectiveness. In: *G. Borich* (Hrsg.): The Appraisal of Teaching: Concepts and Process. Reading, Mass. 1977.

Borich, G., Calkins, D., Malitz, D., Erlich, O., Kugle, C. & Pascone, M.: Classroom observation data: A compendium of methodological papers. The University of Texas at Austin 1977.

Brophy, J., Evertson, C., Crawford, J., King, C. & Senior, K.: Stability of measures of classroom process behaviors across three different contexts in second and third grade classrooms. The University of Texas at Austin 1975.

Brophy, J. & Evertson, C.: Texas Teacher Effectiveness Study: Classroom Coding Manual. The University of Texas at Austin 1976.

Campbell, J.: A longitudinal study in the stability of teachers' verbal behavior. Science Education, 1972. (56). 89–96.

Cooperative Educational Laboratory: CERLI Verbal-Behavior Classification System. Northfield, Illinois, Cooperative Educational Laboratory 1969.

Crawford, J., Brophy, J. & Evertson, C.: Texas Teacher Effectiveness Project: Stability correlations between first and second year data. The University of Texas at Austin 1975.

Cronbach, L., Gleser, G., Nada, H. & Rajaratnam, N.: The dependability of behavioral measurement: The theory of generalizability for scores and profiles. New York 1972.

Denham, C. & Liebermann, A. (Hrsg.): Time to learn. Washington D. C. 1980.

Doyle, W.: Paradigms for research on teacher effectiveness. In: *Shulman, L. S.* (Hrsg.): Review of Research in Education (Vol. 5). Itasca, Ill. 1978. 69–74.

Emmer, E.: Appendix B: Classroom observation scales. In: *Evertson, C.* et al.: Texas Junior High School Study: Final Report. The University of Texas at Austin 1978.

Erlich, O., Borich, G. & Bomberg, B.: Occurence and generalizability of classroom interactions during reading instruction. The University of Texas at Austin 1978.

Erlich, O. & Borich, G.: Occurence and generalizability of scores on a classroom interaction instrument. Journal of Educational Measurement, 1979 (16). 11–18.

Evertson, C., Anderson, L., Edgar, D., Minter, M. & Brophy, J.: Investigations of stability in junior high Math and English classes: The Texas Junior High School Study. The University of Texas at Austin 1977.

Flanders, N.: Analyzing Teaching Behavior. Reading Mass. 1971.

Gage, N. L.: Paradigms for research on teaching. In: *Gage, N. L.* (Hrsg.): Handbook of research on teaching. Chicago 1963. 94–141.

Gage, N. L.: The scientific basis of the art of teaching. New York & London 1978.

Hanke, B., Mandl, H. & Prell, S.: Soziale Interaktion im Unterricht. München 1973.

Klinzing, H. G. & Bar-On, E.: The structure of interrelations among classroom activities for different instructions before and after training. Paper presented at the annual meeting of the American Educational Research Association. Washington D. C 1975. (ERIC ED 109 233).

Merkens, H. & Seiler, H.: Interaktionsanalyse. Stuttgart et al. 1978.

Moon, T. C.: A study of verbal behavior patterns in primary grade classrooms during science activities. Journal of Research in Science Teaching, 1971 (8), 171–177.

Moon, T. C.: A study of verbal behavior patterns in primary grade classrooms during science activities. Ph. D. dissertation, Michigan State University, East Lansing, Michigan 1969.

Schulz, W. et al. (Bearb.): Verhalten im Unterricht. Seine Erfassung durch Beobachtungsverfahren. In: Ingenkamp (Hrsg.): Handbuch der Unterrichtsforschung. Teil I. Weinheim et al. 1970. Sp. 633–852.

Shavelson, R. & Atwood, N.: Generalization of measures of teaching process. In: Borich, G. (Hrsg.): The Appraisal of Teaching: Concepts and Process. Reading Mass. 1977, 351–387.

Soar, R. S.: Follow Through classroom process measurement and pupil growth (1970–1971). Final report. University of Florida, Gainesville 1975. (ERIC ED 106 297).

Spaulding, R.: The Spaulding teacher activity rating schedule (STARS). Duke University Durham, North Carolina 1967.

Withall, J., Newell, J. & Lewis, W.: Development of classroom observational categories within a communication model. University of Wisconsin 1961.

Verfasser: Prof. Dr. Gary D. Borich. Instructional Systems Lab. College of Education 518. University of Texas at Austin. Austin, TX 78712. USA.

Prof. Dr. Hans Gerhard Klinzing. Universität Tübingen. Zentrum für Neue Lernverfahren. Münzgasse 11. D-7400 Tübingen.