

Differentielle Effekte von Lehrerkommentaren zu Noten bei Schülern¹⁾

Günter Krampen²⁾

Universität Trier

Die Ergebnisse einer feldexperimentellen Prüfung der Effekte verschiedener Arten schriftlicher Lehrerkommentare zu Noten in Mathematikarbeiten werden vorgelegt. Unter Bezug auf Kritikpunkte an den vorliegenden Studien werden spezifische Hypothesen über die differentielle Wirksamkeit dreier Kommentartypen (interindividuell, sachlich und intraindividuell orientiert) abgeleitet und empirisch überprüft. An der Untersuchung waren 385 Schüler aus 13 Realschulklassen beteiligt, die nach Zufall den drei experimentellen Bedingungen und einer Kontrollgruppe zugeordnet wurden. Die Untersuchung erstreckte sich auf ein Schuljahr, in dessen erster Hälfte alle Mathematikarbeiten in den Experimentalgruppen kommentiert wurden. Neben einer Voruntersuchung und Nacherhebungen am Ende der Schulhalbjahre wurden kurze Zwischenerhebungen bei der Rückgabe der Arbeiten durchgeführt. Zentrales Ergebnis ist, daß Interaktionen zwischen der Kommentartart und dem Leistungsstand der Schüler nachgewiesen werden konnten. Entsprechende Effekte zeigen sich für die Noten, für situations-, fach- und bereichsspezifische kognitiv-motivationale Variablen. Nachwirkungen der Kommentierung konnten nur bei den bereichsspezifischen Variablen festgestellt werden. Die Diskussion bezieht sich auf die Validität der Befunde, ihre Bedeutung für die schulische Praxis und ihre entwicklungspsychologischen Implikationen.

Während man heute in der Pädagogischen Psychologie versucht, durch relativ umfangreiche Motivförderungsprogramme das Lern- und Leistungsverhalten von Schülern zu optimieren (siehe im Überblick Krug, 1983), beschritt man früher scheinbar einfachere Wege. Klassisch ist dafür etwa die Studie von Page (1958) zu nennen, deren Ergebnisse in eine Vielzahl von Lehr- und Einführungsbüchern zur Pädagogischen Psychologie (z. B. Mathis, Cotton & Sechrest, 1970; Gage & Berliner, 1977; Schmidt, 1978) und zur Methodologie (z. B. Campbell & Stanley,

1) Die Studie wurde von der Deutschen Forschungsgemeinschaft (DFG) unter dem Aktenzeichen Kr 788 gefördert.

2) Für kritische und konstruktive Kommentare zu einer ersten Version dieser Arbeit danke ich der Kollegin Frau B. Rollett und den Herrn Kollegen P. Aymanns, J. Funke, R. Grobe, M. Hofer, A. von Eye und H. Wottawa ganz herzlich.

1963) Eingang gefunden haben und die z. T. auch zu praktischen Empfehlungen oder Vorschriften für Lehrer zur Kommentierung von Zensuren in Klassenarbeiten geführt haben. Page (1958) hat festgestellt, daß Schüler, deren Note in einer schriftlichen Prüfungsarbeit in standardisierter oder freier Form kommentiert wurde, in einem späteren Leistungstest besser abschnitten. Leider konnten die Befunde von Page in einer Reihe von Replikationsversuchen (mit z. T. differenzierterem Vorgehen; s. u.) nicht durchgängig bestätigt werden. So führte auch die metaanalytische Betrachtung von 13 Replikationsstudien zu der Originalarbeit von Page (1958), die Stewart & White (1976) vorlegten, zu einem statistisch nicht signifikanten Ergebnis. Die Autorinnen ziehen den Schluß, daß bislang keine konsistenten Befunde zur Wirkung schriftlicher Lehrerkommentare auf das Leistungsverhalten vorliegen, und daß man allenfalls vermuten kann, (1) daß Kommentare bei älteren Schülern am ehesten Konsequenzen zeigen, (2) daß die Kommentarinhalte eine Rolle spielen und (3) daß längerfristige Wirkungen (im Sinne einer tatsächlichen Steigerung von Leistungsmotiven) kaum zu beobachten sind. Ähnlich skeptisch hat sich unlängst auch Krug (1983) nach einem Literaturüberblick über die Wirkung umfassender Motivförderungsprogramme bei Schülern geäußert. Durch eine engere theoretische Anbindung solcher Förderungsprogramme (an die kognitivistische Motivationstheorie vom Erwartungs-Wert-Typ; im Überblick Heckhausen, 1980, S. 172ff. und S. 564ff.; Krampen, 1982a, S. 25ff.) hofft Krug (1983) zu einer präziseren Identifikation relevanter Variablen und zu besseren Vorhersagen über die Programmwirkungen zu kommen. Ähnliches soll hier für die Evaluation der Effekte schriftlicher Lehrerkommentare zu Noten in Prüfungsarbeiten versucht werden. Mehr als zwei Dekaden nach der klassischen Studie von Page (1958) soll diese Fragestellung also mit einer expliziten theoretischen Anbindung, die zugleich eine veränderte Methodologie (auf Erhebungs- und Auswertungsebene) impliziert, erneut aufgegriffen werden. Aus der Kritik an Pages Studie, die (1.) die Auswahl der abhängigen Variablen und die zeitliche Erstreckung der Untersuchung, (2.) die mangelhafte Spezifikation der unabhängigen Variablen und (3.) die Untersuchungsmethodik betrifft, werden unter Bezug auf die kognitivistische Motivationstheorie und die Theorie der Bezugsnormorientierungen Hypothesen über die Wirkung von Lehrerkommentaren abgeleitet und empirisch überprüft. Im Vergleich zu den umfassenden Motivförderungsprogrammen, die an unterschiedlichsten Bedingungsvariablen des Lern- und Leistungshandelns von Schülern ansetzen (siehe etwa Rheinberg, 1980; Krug, 1983), muß aber beachtet werden, daß hier nur die Effekte eines einzigen situativen Faktors (die mehr oder weniger ausführliche Rückmeldung von Leistungsergebnissen) untersucht werden. Daraus ergibt sich a priori die Erwartung, daß nur kleine, allenfalls mittlere Effektstärken aufgedeckt werden können.

Abhängige Variablen und Dauer der Untersuchung. Ohne weiteren theoretischen Bezug beschränkte sich Page (1958) auf die Erhebung von Zensuren als

abhängige Variable. Neben der in Zweifel stehenden Objektivität und Reliabilität von Zensuren muß dabei auch die Fragen nach der potentiellen Variabilität dieser abhängigen Variablen beachtet werden. Gerade dann, wenn nur sehr kurze Zeiträume untersucht werden (Page beschränkte sich ebenso wie die meisten anderen Autoren auf die Zeitspanne von drei bis vier Wochen zwischen zwei Prüfungsarbeiten), stellt sich die Frage, ob die abhängige Variable in dieser Zeit von den Schülern überhaupt verändert werden kann, ob nicht vielmehr eventuell Lernrückstände, Fähigkeitsmängel oder auch ein sehr guter Leistungsstand, bei dem Verbesserungen mit der Notenskala nicht mehr abgebildet werden können, der Variabilität dieser abhängigen Variablen von vornherein entgegenstehen. Daraus ergeben sich zwei Konsequenzen. Zunächst sollten Untersuchungen längerfristig angelegt sein, um zeitlich-kumulative und langfristige Effekte der Kommentierung erfassen zu können. Sinnvoll ist dabei auch das Absetzen der Kommentierung nach einer längeren Kommentierungszeit, um die Frage zu klären, ob Nachwirkungen zu beobachten sind. Die zweite Konsequenz betrifft die Auswahl der abhängigen Variablen. Da in den meisten Replikationsstudien (siehe etwa Fittkau & Langer, 1974; Stewart & White, 1976) der Nachweis kurzfristiger Effekte schriftlicher Kommentare auf Schulleistungen mißlang, liegt die Vermutung nahe, daß sich Wirkungen nicht nur erst mit der Zeit, sondern dabei auch über die Vermittlung kognitiv-motivationaler Variablen zeigen. Dies bedeutet, daß nicht nur die Zeitspanne der Untersuchung ausgedehnt werden sollte, sondern daß neben Leistungsvariablen auch solche zwischen situativen Faktoren und dem Leistungshandeln von Schülern vermittelnden Variablen erhoben werden sollten. Die kognitivistische Motivationstheorie liefert einen konzeptuellen Rahmen für die Variablenauswahl. Situations- und handlungsspezifische Variablen wie etwa die Zufriedenheit mit der Note, subjektive Folgebewertungen, Handlungs-Ergebnis/Folge-Erwartungen und Selbstkonzepte eigener Fähigkeiten haben sich bislang in Analysen und Prognosen des Leistungshandelns von Schülern recht gut bewährt (siehe etwa Krampen, 1979; Heckhausen & Rheinberg, 1980; Krampen & Lehmann, 1981; Grobe & Hofer, 1983). Daneben ist an bereichsspezifische Variablen zu denken, die als Generalisierungen dieser situationsspezifischen Variablen mit Erwartungs-Wert-Modellen in Beziehung stehen, sich jedoch nicht mehr auf bestimmte Situationen und Handlungen, sondern auf bestimmte Situations- und Handlungsklassen (hier etwa Schulleistungen allgemein, Schuleinstellungen) beziehen. Solche situationsübergreifenden Variablen, die erwartungs-wert-theoretisch rekonstruierbar sind liegen mit den Konzepten der Kontrollüberzeugungen (vgl. Krampen, 1982a, 1984a) und der Prüfungsangst (vgl. Helmke, 1983; Pekrun, 1983, 1984) vor.

Fittkau & Langer (1974) konnten kurzfristige (positive) Effekte allgemein ermutigender Lehrerkommentare auf die Prüfungsangst und Schulunlust von Schülern feststellen, Butterworth & Michael (1975) gelang analoges für leistungsbezogene Kontrollüberzeugungen und Schuleinstellungen. Kurz seien die entwicklungspsychologischen Implikationen dieser Überlegungen erwähnt. Bedingungs-wissen aus prospektiv angelegten Interventionsstudien ist ein wesentliches Fundament einer ange-

wandten Entwicklungspsychologie (siehe Montada & Schmitt, 1982). Gelingt der Nachweis relevanter Effekte der Lehrerkomentierung auf diese allgemeineren Variablen mit persönlichkeitspsychologischem Status, so wird gleichzeitig ein Beitrag zu diesem Wissen geliefert.

Zusammenfassend ergibt sich aus diesen Überlegungen für die geplante Untersuchung, daß als abhängige Variablen neben den Schulleistungen situations-spezifische kognitiv-motivationale, schulleistungsspezifische und schulspezifische Variablen erhoben werden sollen. Liegt bei den drei zuletzt genannten Variablenklassen bereits eine potentiell höhere inter- und intraindividuelle Variabilität vor, so bleibt zu beachten, daß für einige Analysen bestimmte Schülergruppen ausgefiltert werden müssen (Schüler mit sehr guten Noten dürfen so etwa nicht in Ergebnisanalysen zu Fragen nach der Verbesserungsmotivation eingehen). Als erste Arbeitshypothesen lassen sich aus dem Gesagten die Erwartungen ableiten, (1.) daß Lehrerkommentare zeitlich kumulativ wirken, (2.) daß sich Kommentare am deutlichsten (wegen der zeitlichen Staffelung der Effekte) in situationsspezifischen kognitiv-motivationalen Variablen, dann in Leistungsvariablen und deren subjektiven Abbildungen (z. B. Zufriedenheit mit der Note) und dann in schulleistungsspezifischen und schulspezifischen Variablen (z. B. leistungsspezifische Kontrollüberzeugungen, Prüfungsangst) niederschlagen, und (3.) daß die Effektdauer der Kommentierung (Nachwirkung) bei den bereichsspezifischen Variablen am höchsten ist, da sie weniger stark situativ gebunden sind.

Spezifikation der Kommentierungsart. Der oben genannte zweite Kritikpunkt betrifft die mangelhafte Spezifikation und Definition der unabhängigen Variable in den vorliegenden Untersuchungen. In Anlehnung an verschiedene Modelle der Leistungsbewertung (siehe etwa Zielinski, 1974; Klauer, 1982) und an die Unterscheidung von Bezugsnormorientierungen (siehe etwa Rheinberg, 1980) können am sozialen Bezugssystem (der Klasse), an kriterialen (sachlich-lehrstoffbezogenen) Normen und am individuellen Bezugssystem (Leistungskurve, Lern- und Leistungsbedingungen des einzelnen Schülers) orientierte Lehrerkommentare unterschieden werden (siehe hierzu ausführlicheres bei Krampen, 1982b, 1984b). In den vorliegenden Studien wurden mit wenigen Ausnahmen Mischtypen dieser Kommentararten verwendet, wobei die gängige Hypothese, daß schriftlich kommentierte Noten im Vergleich zu nicht kommentierten positive Effekte bei den Schülern zeigen, nicht durchgängig bestätigt werden konnte. Als eine erste Differenzierung wird daher hier die Hypothese formuliert, daß sich nicht nur die Effekte kommentierter und nicht kommentierter Noten unterscheiden, sondern daß sich auch bei inhaltlich verschiedenartigen Kommentararten unterschiedliche Auswirkungen in kognitiv-motivationalen Variablen ergeben (Arbeitshypothese 4). Eine zweite Differenzierung, die diese Haupteffekthypothese durch Hinzunahme einer weiteren unabhängigen Variable spezifiziert und somit zu einer Interaktionshypothese führt, wird im folgenden vorgenommen.

Bei der verbalen, mehr oder weniger festgelegten Umschreibung der Ziffernnote steht eine soziale Bezugsnormorientierung des Lehrers im Vordergrund, für die Rheinberg (1980) allgemein negative Effekte bei leistungsschwächeren und eher positive oder neutrale Effekte bei leistungsstärkeren Schülern berichtet. Es ist also zu vermuten, daß die Wirkung dieses Kommentartyps mit dem Leistungsstand des Schülers interagiert, was in vorliegenden Arbeiten (siehe etwa Butterworth & Michael, 1975; im Überblick Stewart & White, 1976) nicht beachtet wurde, so daß deren negative Befunde eventuell durch Kompensationseffekte erklärt werden können.

Der an sachlichen Gütemaßstäben orientierte Kommentar gibt allen Schülern gezielte Hinweise auf ihren sachbezogenen Kenntnis- und Leistungsstand, woraus zunächst die Hypothese resultiert, daß er bei allen Schülern positiv wirkt. In einer Voruntersuchung konnten kurzfristige Effekte dieser bis dahin nicht untersuchten Kommentierungsart auf kognitiv-motivationale Variablen (Zufriedenheit mit der Note, Kausalattributionen der Note), nicht jedoch auf Leistungsvariablen festgestellt werden (siehe Krampen, 1982b). Es muß aber bedacht werden, daß bei einer Überforderung des Schülers oder bei zu großen Lerndefiziten mehr oder weniger stereotype negative Rückmeldungen gehäuft auftreten können, die positive Effekte unwahrscheinlich machen. Auch hier wird somit eine (schwache, da nur Extremfälle betreffende) Interaktion zwischen Kommentarart und Leistungsstand der Schüler erwartet.

Bei dem an intraindividuellen Leistungsvergleichen orientierten und die spezifische Lern- und Leistungssituation beachtenden Kommentar ist aufgrund der in ihm verwirklichten individuellen Bezugsnormorientierung mit positiven Effekten bei allen Schülern zu rechnen (siehe Rheinberg, Kühmel & Duscha, 1979; Rheinberg, 1980). Dies mag bei entsprechender Anregung und Förderung (und nicht nur stereotypem Loben; vgl. Tacke & Linder, 1981) auch für leistungsstarke Schüler gelten, wobei aber zu beachten ist, daß sich dies weniger in Zensuren, subjektiven Folgebewertungen und Handlungs-Ergebnis-Erwartungen (aufgrund ihrer mangelnden Variabilität) als in anderen Variablen (z. B. Interesse am Fach, Kontrollüberzeugungen) niederschlagen wird. Im Gegensatz zu den anderen Kommentartypen setzt diese Art der Kommentierung beim Lehrer eine vertiefte Kenntnis des Schülers voraus. Butterworth & Michael (1975) konnten zeigen, daß solche individuum-zentrierten Kommentare sowohl in späteren Prüfungsleistungen als auch in Maßen zu Schuleinstellungen und Kontrollüberzeugungen positive Effekte zeigten, die für am sozialen Bezugssystem orientierte Kommentare nicht nachgewiesen werden konnten; Interaktionshypothesen wurden allerdings nicht aufgestellt und überprüft.

Bei der bislang in der Literatur anzutreffenden Verwendung unspezifischer Kommentare ohne Beachtung des ipsativen Leistungsstandes der Schüler ist also zu vermuten, daß wertvolle Informationen über die *differentielle Wirksamkeit* von Kommentaren verschenkt werden, die in der Pädagogischen Psychologie etwa unter das Schlagwort der *treatment-aptitude interaction*, in der Klinischen Psychologie unter das der differentiellen Indikation fallen. Zielsetzung soll daher hier die Beantwortung der Frage sein, von welcher Kommentierungsart bei welchen Schülern die motivational positivsten Effekte erwartet werden können. Nach Arbeitshypothese (5) werden daher Interaktionen zwischen dem Leistungsstand der Schüler und der Kommentarart erwartet (wie sie oben beschrieben wurden). Mit dieser Interaktionshypothese (5) wird somit die (Haupteffekt-) Hypothese (4) weiter ausdifferenziert.

Erwartungseffekte. Der dritte der genannten Kritikpunkte an den bisherigen Arbeiten zur Wirkung von Kommentaren betrifft die Untersuchungsmethodik. Neben der notwendigen Erweiterung des Erhebungsinstrumentariums (s. o.) ist hier vor allem die zumeist nur ungenügende Kontrolle von Erwartungseffekten und die Art der statistischen Hypothesenprüfung (siehe dazu unter „Methode“) gemeint (vgl. auch Krampen, 1982b, 1984b).

Wurden Erwartungseffekte auf der Schülerseite meist dadurch kontrolliert, daß die beteiligten Schüler nicht über den Sinn und Zweck der Untersuchung informiert werden („Blindversuche“, bei denen allerdings durchaus Abweichungen des Unterrichtsalldages vom Gewöhnlichen noch Einfluß haben können; siehe Page, 1958; Rheinberg et al., 1979), so wurden Erwartungseffekte auf der Lehrerseite kaum beachtet. Man muß hier gar nicht an die mögliche Wirkung gutgläubiger Erwartungen von Lehrern im Sinne des Pygmalioneffektes denken, es reicht aus, den aus der Experimentalpsychologie hinreichend bekannten Versuchsleiter-Effekt heranzuziehen. Immer dann, wenn ein Lehrer dazu aufgefordert ist, bei einigen (nach externen Kriterien bestimmten) Schülern einer Klasse auf eine bestimmte Art und Weise zu reagieren (etwa Noten zu kommentieren), bei anderen jedoch nicht (siehe etwa Page, 1958; Fittkau & Langer, 1974; Rheinberg et al., 1979; Tacke & Linder, 1981), muß mit solchen Erwartungseffekten gerechnet werden. Dies gilt auch für den Fall, in dem der Lehrer bei zwei von ihm unterrichteten Klassen unterschiedliche Kommentierungsarten einsetzen soll. Problematisch ist dabei, daß der Lehrer weiß, welcher Schüler welche Art der Kommentierung (oder keinen Kommentar) erhält, was vor allem dann, wenn der Lehrer subjektive Theorien über die Kommentarwirkung hat, verzerrend auf sein Unterrichts- und evtl. auch auf sein Leistungsbeurteilungshandeln wirken kann. Dann, wenn auch individuell orientierte Kommentare eingesetzt werden sollen, scheidet die in der Experimentalpsychologie gängige Lösung dieses Problems (Doppelblindversuche) aus. Ein Ausweg besteht darin, daß jeder Lehrer jeweils nur eine Versuchsbedingung realisieren muß. Auftretende Erwartungseffekte können so zumindest nicht mehr differentiell für verschiedene experimentelle Gruppen wirken. Zentral ist bei diesem Vorgehen allerdings die Vergleichbarkeit der verschiedenen experimentellen Gruppen (die aus jeweils mehreren Klassen bestehen sollten) und der an der Untersuchung beteiligten Lehrer. Im Randomisierungsprozeß ist somit die Schulklasse (als präexistierende Gruppe) die Einheit: Experimental- und Kontrollgruppen bestehen aus mehreren ungeteilten Klassen, wobei die Zuordnung der Klassen (und ihrer Lehrer) zu den Versuchsbedingungen nach Zufall erfolgt. Die auswertungstechnischen Implikationen dieses Vorgehens (siehe etwa Lindquist, 1953; Campbell & Stanley, 1963) rückten in der deutschen Pädagogischen Psychologie bei der kritischen Kommentierung von Evaluationsstudien zu Schulversuchen in den Blickpunkt. Die Kommentatoren (Wottawa, 1981; Krauth, 1983) dieser Studien machen darauf aufmerksam, daß häufig das Individuum als Analyseeinheit gewählt wird, obwohl weder die Schüler noch die Schulklassen den verschiedenen „Versuchsbedingungen“ (dort: Schultypen) nach Zufall zugeordnet wurden. In der vorliegenden Arbeit besteht diese Problematik nur bei der Wahl des Individuums als Analyseeinheit, nicht bei der der Klasse als Analyseeinheit, da diese zufallsmäßig den Experimentalbedingungen zugeordnet werden. Analysen auf Individuum-Ebene werden aber dann abgesicherter möglich, wenn die von Cook & Campbell (1979; siehe auch Krauth, 1983) vorgeschlagene Strategie eingesetzt wird: Kann die Vergleichbarkeit der Individuen aus verschiedenen Klassen mit Bezug auf Variablen, die als potentielle Alternativerklärungen möglicher Effekte in Frage kommen, gesichert werden, so kann es gelingen zu einer „Interpretation der Daten zu kommen, die gegen alle offensichtlichen Einwände abgesichert ist“ (Krauth, 1983, S. 5). Solche Variablen sind im vorliegenden Kontext zunächst alle als abhängige Variablen verwendeten Größen, die Leistungsorientierung in den Klassen und das Klassenklima sowie die Bezugsnormorientierungen der Lehrer vor dem Beginn der Experimentalphase.

Um Erwartungseffekte auf der Schülerseite weitgehend ausschließen zu können, soll die gesamte Untersuchung in den normalen Ablauf des Mathematikunterrichts in einem ganzen Schuljahr integriert sein. Das Fach Mathematik wurde vor allem deswegen ausgewählt, weil dafür in Rheinland-Pfalz keine Anweisungen zur Notenkomentierung in Prüfungsarbeiten für Lehrer bestehen (im Gegensatz zu Deutscharbeiten), weil mit geringeren Objektivitätsproblemen bei der Notengebung in schriftlichen Arbeiten gerechnet werden kann und weil der Lehrstoff für die Klassenstufen eines Schultyps in diesem Fach in höherem Maße standardisiert ist als in anderen Schulfächern.

Methoden

Versuchsplan und Versuchspersonen. In Kooperation mit dem Schulreferat der Bezirksregierung Trier wurden im Sommer 1982 alle Realschulen des Regierungsbezirkes Trier und der Stadt Trier mit der Bitte um Kooperation angeschrieben³⁾. Die Mathematiklehrer der Schulen wurden nach ihrer Bereitschaft zur honorierten Mitarbeit an der Untersuchung gefragt und um die Rücksendung eines kurzen Fragebogens (unterrichtete Klassenstufen, Art der Notengebung etc.) gebeten. Aus den Rückmeldungen wurden unter Beachtung einer annähernden Gleichverteilung verschiedener Klassenstufen insgesamt 13 Realschulklassen und deren Fachlehrer in Mathematik für die Untersuchung nach Zufall ausgewählt. Aus den Rückmeldungen ergab sich ferner, daß keiner der Lehrer Noten in Prüfungsarbeiten durch schriftliche Kommentare ergänzte. Die Stichprobe der an der Untersuchung beteiligten Schüler beträgt insgesamt $N = 385$ (190 Mädchen und 195 Jungen; $\bar{x} = 13.9$, $s = 1.40$ Jahre; Klassenstufe 6: 1x, 7: 2x, 8: 2x, 9: 4x und 10: 4x; Klassenstärke: $24 \leq N \leq 34$). Alle Schüler waren mit der Teilnahme an der Untersuchung, die ihnen bei der ersten Erhebung kurz (aber ohne jeden Hinweis auf das experimentelle Vorgehen) in ihrer Organisation erläutert wurde, einverstanden.

Die 13 Realschulklassen wurden nach dem Zufall (wobei die Klassenstufe kontrolliert wurde, um eine annähernde Gleichverteilung der Klassenstufen auf die Gruppen zu gewährleisten) den vier folgenden Gruppen zugeordnet: (1) *Experimentalgruppe I*: Am sozialen Vergleich in der jeweiligen Klasse orientierte Kommentierung aller Mathematikarbeiten im ersten Schulhalbjahr. Diese Gruppe besteht aus $N = 87$ Schülern aus drei Klassen der Stufe 6, 9 und 10; (2) *Experimentalgruppe II*: An einem sachlichen Gütemaßstab orientierte Kommentierung aller Mathematikarbeiten im ersten Schulhalbjahr ($N = 116$ Schüler aus drei Klassen der Stufe 8, 9 und 10); (3) *Experimentalgruppe III*: Individuell orientierte Kommentierung aller Mathematikarbeiten im ersten Schulhalbjahr ($N = 92$ Schüler aus vier Klassen der Stufe 7, 8, 9 und 10); (4) *Kontrollgruppe (Gruppe IV)*: Keine schriftliche Kommentierung der Noten in den Mathematikarbeiten ($N = 90$ Schüler aus drei Klassen der Stufe 7, 9 und 10). Die Fachlehrer wurden in Einzelgesprächen in die Art der von ihnen verlangten Kommentierung der Zensuren eingewiesen. Jeder Lehrer hatte nur einen Typus der Kommentierung (bzw. keine) einzusetzen. Merkblätter erläutern zusätzlich die Art des verlangten Vorgehens (siehe Krampen, 1984b)⁴⁾. Bei den Zwischentestungen (jeweils nach der Rückgabe von Mathematikarbeiten) wurde die Art der Kommentierung stichprobenhaft kontrolliert; Fragen und Probleme der Lehrer wurden besprochen und ausgeräumt. Jeder Lehrer war nur über die Art des von ihm zu vergebenden Kommentars informiert. Nach dem Erhalt der Zwischenzeugnisse (Februar 1983) wurde die Kommentierung der Noten in den Experimentalgruppen abgesetzt, die Datenerhebungen (s. u.) wurden jedoch weitergeführt. Die durch die Kommentierung der Noten in der ersten Untersuchungsphase entstehende Mehrarbeit der Lehrer wurde aus Mitteln einer Sachbeihilfe der DFG honoriert.

Neben einer umfangreichen Voruntersuchung (t_1 ; zu Beginn des Schuljahres 1982/83 im September 1982), einer umfangreichen, die erste Untersuchungsphase abschließenden ersten Nachuntersuchung (t_5 ; unmittelbar nach Ausgabe der Halbjahreszeugnisse im Februar 1983) und einer umfangreichen, die zweite Untersuchungsphase abschließenden zweiten Nachuntersuchung (t_9 ; nach Erhalt der Versetzungs- bzw. Abschlußzeugnisse im Juli 1983) wurden jeweils unmittelbar nach der Rückgabe von Mathematikarbeiten kurze Zwischenuntersuchungen ($t_2, t_3, t_4, t_6, t_7, t_8$;

3) Für die Unterstützung bei der Rekrutierung der Untersuchungsstichproben danke ich Herrn Reg.-Direktor Wacker (Bezirksregierung Trier) ganz herzlich. Mein Dank gilt natürlich auch allen an der Untersuchung beteiligten Lehrern und Schülern für ihre Kooperationsbereitschaft.

4) Der Arbeitsbericht, in dem die genauen Instruktionen für die Art der Kommentierungen beschrieben sind, kann vom Autor angefordert werden.

z. B.: t_3 = Erhebung nach der Rückgabe der zweiten Mathematikarbeit) durchgeführt. Die Termine dieser Zwischenuntersuchungen variieren zeitlich zwischen den Klassen, da die Prüfungsarbeiten in den normalen Unterrichtsablauf integriert waren (es handelt sich um die normalen, für die Zeugnisnoten relevanten Mathematikarbeiten). Alle Datenerhebungen wurden anonym durch Versuchsleiter durchgeführt, die darauf achteten, daß im Falle der Lehreranwesenheit der Lehrer keinen Einfluß auf die Schüler bei der Fragebogenbearbeitung nahm. Die Zuordnung der Fragebogen von verschiedenen Erhebungszeitpunkten zu den Schülern erfolgte durch einen stabilen Personen-Code. In Abbildung 1 ist der realisierte feldexperimentelle Versuchsplan in der üblichen, von Campbell & Stanley (1963) übernommenen Notation dargestellt.

R	O_1	X_a	O_2	X_a	O_3	X_a	O_4	O_5	O_6	O_7	O_8	O_9
R	O_1	X_b	O_2	X_b	O_3	X_b	O_4	O_5	O_6	O_7	O_8	O_9
R	O_1	X_c	O_2	X_c	O_3	X_c	O_4	O_5	O_6	O_7	O_8	O_9
R	O_1		O_2		O_3		O_4	O_5	O_6	O_7	O_8	O_9

Abb. 1. Versuchsplan des Feldexperiments

(R = Randomisierung, O_x = Meßzeitpunkt t_x , X_y = Versuchsbedingung / Kommentarart y).

Erhebungsmethoden. Um die Bezugsnormorientierung der Lehrer, die mit der unabhängigen Variablen der Kommentierungsart konfundiert sein kann, zu kontrollieren, bearbeiteten alle beteiligten Lehrer vor Beginn der Experimentalphase den von Rheinberg (1980, S. 114ff.) vorgelegten „Fragebogen zur Erfassung von Bezugsnorm-Orientierungen“ (FEBO), der aus insgesamt 39 Items besteht. Für die fünf Inhaltsbereiche „Leistungsvergleich“, „Kausalattribution“, „Individualisierungstendenz“, „Erwartung“ und „Sanktionierungsstrategie“ werden mit dem FEBO interindividuelle Unterschiede in der Bezugsnormorientierung auf der Dimension soziale versus individuelle Bezugsnorm abgebildet.

Bei der *Voruntersuchung* (t_1) bearbeiteten die Schüler das u. a. die folgenden Teile umfassende Erhebungsinstrumentarium: (1) *Lernsituationstest* (LST) von Kahl (1977; siehe auch Kahl, Bachmann & Witte, 1977) zur Erfassung der subjektiven Wahrnehmung der unterrichtlichen Lernsituation durch die Schüler (Klassenklima). Der LST umfaßt mit 65 Items die folgenden faktorenanalytisch gewonnenen Skalen: (a) wahrgenommene Kohäsion in der Klasse, (b) Identifikation mit der Unterrichtsarbeit, (c) wahrgenommene Leistungsanforderungen im Unterricht und (d) wahrgenommene Betonung von Wettbewerb und Ordnung in der Klasse. Die Testhalbierungsreliabilität (Spearman-Brown) der LST-Subskalen liegt in der vorliegenden Stichprobe bei $r_{tt} \geq .78$. (2) *Fragebogen KOM I* zur Erfassung des objektiven und subjektiv wahrgenommenen Leistungsstands in Mathematik sowie von situations- und fachspezifischen kognitiv-motivationalen Variablen. U. a. wurden die folgenden Daten erhoben: Mathematiknote des letzten Zeugnisses, Zufriedenheit damit (7-stufige graphische Schätzskala), drei Selbsteinschätzungen eigener mathematischer Fähigkeiten (sozialer, intraindividueller und sachlicher Aspekt; 6-stufige Schätzskalen), Verbesserungsmotivation (6-stufige Schätzskala), Valenz einer Verbesserung, Verbesserungserwartung, mathematisches Interesse (jeweils 7-stufige bipolare Schätzskalen) und erlebte Überforderung im Mathematikunterricht (ja/nein). Interdependanzanalysen dieser Variablen zeigen, daß es gelungen ist, diese Variablen intern konsistent und reliabel zu erfassen (siehe Krampen, 1984b). (3) *IPC-PL-Fragebogen* zur Erfassung von drei Aspekten von Kontrollüberzeugungen im Bereich des Problemlösehandelns. Mit diesem in Anlehnung an die neuere Literatur zur Differenzierung des Konstrukts der Kontrollüberzeugungen (vgl. Krampen, 1982a) neu konstruierten Instrument werden über 24 Items die folgenden Variablen erhoben: (a) Internalität,

(b) sozial bedingte Externalität und (c) fatalistische Externalität bezüglich des Problemlösehandelns in intellektuellen Realitätsbereichen. Erste Befunde zu seiner Reliabilität (u. a. $r_{tt} \geq .55$) und Validität bestätigen die Tauglichkeit des IPC-PL für Gruppenvergleiche (siehe hierzu Krampen, 1984b). (4) *Angstfragebogen für Schüler* (AFS) von Wiczerkowski et al. (1975) zur Erfassung von (a) manifester Angst, (b) Prüfungsangst, (c) Schulunlust und (d) sozialer Erwünschtheit (in der vorliegenden Stichprobe u. a.: $r_{tt} \geq .72$).

Bei den *Zwischenuntersuchungen* (t_2 bis t_4 und t_6 bis t_8), die jeweils unmittelbar nach der Rückgabe einer Mathematikarbeit durchgeführt wurden, wurden verschiedene Varianten des *Fragebogens KOM II* eingesetzt. Neben der Note der zurückerhaltenen Arbeit wurden mit ihm verschiedene kognitiv-motivationale Variablen (wie Zufriedenheit mit der Note, erlebte Urteilsgerechtigkeit etc.) erhoben.

Bei den beiden *Nachuntersuchungen*, die die erste (t_5) und die zweite Untersuchungsphase (t_9) abschlossen, wurde mit Ausnahme des LST zur Erfassung des Klassenklimas das gleiche Erhebungsinstrumentarium wie bei der Voruntersuchung eingesetzt (s. o.). Lediglich bei der zweiten Nacherhebung nach den Versetzungszeugnissen (t_9) wurde der Fragebogen KOM I um einige Items (z. B. Verbesserungsmotivation) reduziert, da sich diese Erhebung bei den Schülern der 10. Klasse auf die Schulabschlußzeugnisse bezog.

Auswertungsmethoden. Im Unterschied zu den älteren Studien zur Wirkung von Kommentaren muß aufgrund der Interaktionshypothesen die statistische Hypothesenprüfung über mehrfaktorielle Varianzanalysen durchgeführt werden. Ein hierarchisch geschachteltes Vorgehen (zunächst Prüfung globaler, dann Prüfung spezifischer Hypothesen durch die Folge von bifaktoriellen multivariaten Varianzanalysen, unifaktoriellen Varianzanalysen und Einzelmittelwertvergleichen nach Duncan) soll gegen die Notwendigkeit von Alpha-Justierungen absichern, die ohnedies dann nicht angebracht sind, wenn die Forschungshypothese nur dann vollständig gestützt ist, wenn alle (bezogen auf die verschiedenen abhängigen Variablen) Alternativhypothesen akzeptiert werden können (siehe Westermann & Hager, 1983). In diesem Fall ist die Aussagekraft jedes einzelnen statistischen Tests wichtiger als ein angepaßtes Signifikanzniveau, das somit global auf $p = .05$ festgelegt wird. Um die Bedeutung der Ergebnisse beurteilen zu können, werden ergänzend zu den Angaben über die statistische Signifikanz Schätzungen für die Effektstärke als Maße der praktischen Signifikanz abgegeben. Im Falle signifikanter Befunde in Varianzanalysen werden Omega²-Werte (bezogen auf die Gesamtvarianz) als Schätzung der durch die unabhängigen Variablen in der abhängigen Variablen erklärten Varianz (siehe Hays, 1963) berechnet, wobei zu ihrer Beurteilung die Kriterien Cohens (1977²) angelegt werden. Wie in der Einführung schon vermerkt wurde, werden aufgrund der Interventionsart, die sich isoliert auf einen einzigen (situativen) Faktor (Art der Leistungsrückmeldung bei Prüfungsarbeiten) bezieht, nur kleine bzw. mittlere Effekte erwartet.

Zum Vorgehen bei der Datenauswertung ist noch zu vermerken, daß zweigleisig verfahren wird. Zunächst wird die Vergleichbarkeit der Individuen in den experimentellen Gruppen überprüft, um mögliche Alternativerklärungen, die sich auf a priori Unterschiede zwischen den Klassen und den Schülern beziehen, von vornherein weitgehend ausschließen zu können. Analoges gilt für die Kontrolle von Klasseneffekten: In Varianzanalysen (mit dem Faktor „Klassenzugehörigkeit“) wird überprüft, ob die Varianz zwischen den Klassen signifikant größer ist als die Varianz innerhalb der Klassen. Ist dies nicht der Fall, so wird zunächst das Individuum als Analyseeinheit gewählt (zu dieser Strategie siehe Cook & Campbell, 1979; Krauth, 1983). Die so erhaltenen Befunde sollen dann in (konservativeren) Berechnungen, in die die Klassenmittelwerte als Analyseeinheit eingehen, abgesichert werden. Schüler, bei denen Variablen fehlen (*missing data*), werden analysespezifisch ausgeschlossen.

Ergebnisse⁵⁾

Zentral für unsere Fragestellung sind die Arbeitshypothesen (4) und (5), die die globale bzw. differentielle Wirkung der Lehrerkommentierung betreffen. Auf sie bezogene Datenanalysen stehen im Vordergrund der Ergebnisdarstellung, wobei auch Aussagen über die Arbeitshypothesen (1) bis (3), die sich auf die zeitlich-kumulative und variablen-spezifische Wirksamkeit der Kommentierung beziehen, gemacht werden können.

Vergleichbarkeit der Untersuchungsgruppen. Die Rohwerte der an der Untersuchung beteiligten Lehrer im „Fragebogen zur Erfassung von Bezugsnorm-Orientierungen“ (FEBO) variieren zwischen 131 und 148 Punkten, was in der von Rheinberg (1980) für eine Stichprobe von $N = 213$ Lehrern berichteten Rohwertverteilung im mittleren Bereich liegt ($\bar{x} + / - 1s$). Bei den an der Untersuchung beteiligten Lehrern liegen somit a priori keine extremen Bezugsnormorientierungen vor; es bestehen zwischen ihnen (bezogen auf die Rohwertverteilung bei Rheinberg, 1980) auch keine großen Unterschiede.

Zur empirischen Überprüfung der Vergleichbarkeit der vier Untersuchungsgruppen wurden (unifaktorielle) varianzanalytische Vergleiche für alle bei der Voruntersuchung erhobenen Variablen durchgeführt (Analyseeinheit: Individuum). Eine multivariate Varianzanalyse zeigt, daß zwischen den Gruppen keine statistisch bedeutsamen Mittelwertsunterschiede in den LST- (Klassenklima), IPC-PL- (Kontrollüberzeugungen) und AFS-Skalen (Angst und Schulunlust) bestehen ($F(30/991) = 0.470$; n. s.). Univariate Varianzanalysen für die einzelnen Variablen deckten auch keine variablen-spezifischen Mittelwertsunterschiede auf ($F(3/364) \leq 1.86$; n. s.). Über Bartlett-Box-Tests konnten keine Hinweise auf Varianzheterogenität festgestellt werden ($F \leq 2.50$). Analoges gilt für die zu t_1 erhobenen kognitiv-motivationalen Variablen und die (letzte) Zeugnisnote in Mathematik (KOM I-Fragebogen). Die multivariate Varianzanalyse ergab einen Wert von $F(27/1066) = 0.214$ (n. s.), die univariaten F-Werte übersteigen nicht den Wert von $F(3/381) = 1.92$ (n. s.) und durch Bartlett-Box-Tests konnten auch hier keine bedeutsamen Varianzunterschiede zwischen den Gruppen aufgedeckt werden ($F \leq 1.93$). Die an der Untersuchung beteiligten Schülergruppen sind somit vor Beginn der Experimentalphase in Leistungsvariablen, kognitiv-motivationalen Variablen, den verschiedenen Aspekten des wahrgenommenen Klassenklimas und den bereichsspezifischen Persönlichkeits-/Einstellungsmerkmalen vergleichbar. Dies wird für die im folgenden als abhängige Variablen benutzten Daten in Tabelle 1 (in den Tabellenzeilen für t_1) spezifisch dokumentiert, wobei es wesentlich ist, daß die Einheit dieser Analysen das Individuum (nicht die Klasse) ist. Auch Klasseneffekte können aufgrund der Ergeb-

5) Die Ergebnistabellen wurden auf Vorschlag der Herausgeber stark gestrafft und zusammengefaßt; ausführliches Tabellenmaterial kann vom Autor angefordert werden.

nisse in den mit dem Faktor „Klassenzugehörigkeit“ berechneten multivariaten Varianzanalysen weitgehend ausgeschlossen werden: Weder für die zu t_1 erhobenen situationsspezifischen kognitiv-motivationalen Variablen ($F(108/2706) = 0.893$; n. s.) noch für die erhobenen Persönlichkeits- und Einstellungsvariablen ($F(120/2832) = 0.937$; n. s.) zeigt sich, daß die Varianz zwischen den Klassen statistisch bedeutsam größer ist als die Varianz innerhalb der 13 untersuchten Schulklassen.

Effekte auf situationsspezifische kognitiv-motivationale Variablen. Da bei den situationsspezifischen kognitiv-motivationalen Variablen (subjektive Wichtigkeit einer Verbesserung der Mathematikzensur, subjektive Verbesserungs-Erwartung und fachspezifische Verbesserungsmotivation) bei leistungsstarken Schülern mit geringer (oder keiner) Variabilität gerechnet werden muß, wurden für die entsprechenden Datenanalysen jeweils die Schüler, die zu t_1 bzw. t_3 über die Note „sehr gut“ oder „gut“ in Mathematik verfügten, aus den Berechnungen ausgeschlossen. Schüler mit den Noten „mangelhaft“ und „ungenügend“ wurden zu einer Leistungsgruppe zusammengefaßt.

Tabelle 1 zeigt, daß sich die Schüler der vier Untersuchungsgruppen zu t_1 (Voruntersuchung) in diesen Variablen nicht bedeutsam unterscheiden (Haupteffekte für die Gruppenzugehörigkeit zu t_1 sind nicht signifikant; s. o.). Für den zweiten varianz-analytischen Faktor (Leistungsstand zu t_1 : Schüler mit der Note „befriedigend“, „ausreichend“ bzw. „mangelhaft / ungenügend“) ergeben sich dagegen signifikante Haupteffekte, die auf wesentliche Interdependenzen zwischen Leistungsstand und situationsspezifischen kognitiv-motivationalen Variablen zu einem Erhebungszeitpunkt verweisen (siehe auch Krampen, 1979). Wichtig ist, daß sich für t_1 keine bedeutsamen Interaktionseffekte zeigen (siehe Tabelle 1).

Tab. 1. Übersicht zu den varianzanalytischen Befunden für die zu t_1 , t_3 und t_9 erhobenen Variablen ($225 \leq N \leq 385$)

abhängige Variable	t	varianzanalytische Effekte					
		Kommentarart		Leistungsstand t ₁		Interaktion	
		F	η^2	F	η^2	F	η^2
situationsspezifische kognitiv-motivationale Variablen							
Wichtigkeit der Verbesserung	t ₁	0.88		18.24*	0.271	0.78	
	t ₃	1.47		13.88*	0.211	0.36	
Verbesserungs-Erwartung	t ₁	1.92		7.33*	0.084	1.21	
	t ₃	11.77*	0.123	6.54*	0.061	4.77*	0.053
Verbesserungs-Motivation	t ₁	1.87		11.40*	0.137	0.93	
	t ₃	10.57*	0.104	6.42*	0.053	3.61*	0.042

Fortsetzung Tabelle 1

		varianzanalytische Effekte					
abhängige Variable	t	Kommentarart		Leistungsstand t ₁		Interaktion	
		F	w ²	F	w ²	F	w ²
fachspezifische Variablen							
Zeugnisnote	t ₁	1.25		—		—	
	t ₅	19.77*	0.081	60.04*	0.338	2.22*	0.021
	t ₉	1.57		29.39*	0.241	0.22	
Zufriedenheit mit der Note	t ₁	1.59		—		—	
	t ₅	7.88*	0.054	12.54*	0.122	2.56*	0.034
	t ₉	1.44		6.51*	0.049	0.14*	
mathematisches Interesse	t ₁	0.51		—		—	
	t ₅	13.44*	0.093	18.51*	0.184	2.76*	0.044
	t ₉	0.97		9.77*	0.064	0.53	
bereichsspezifische Persönlichkeitsvariablen							
Internalität	t ₁	0.10		9.97*	0.130	0.42	
	t ₅	3.82*	0.041	7.18*	0.092	2.91*	0.024
	t ₉	2.71*	0.022	5.79*	0.054	1.44	
sozial-bedingte Externalität	t ₁	2.39		11.89*	0.153	0.11	
	t ₅	4.66*	0.051	7.15*	0.073	3.17*	0.034
	t ₉	2.90*	0.032	3.99*	0.031	1.02	
fatalistische Externalität	t ₁	2.39		2.93*	0.023	0.32	
	t ₅	2.71*	0.031	1.82		0.98	
	t ₉	2.45		2.58		0.88	
Schulunlust	t ₁	0.87		4.71*	0.052	0.03	
	t ₅	4.05*	0.063	2.54*	0.031	2.99*	0.042
	t ₉	2.93*	0.031	1.07		0.81	
Prüfungsangst	t ₁	0.44		8.41*	0.106	0.13	
	t ₅	2.75*	0.023	6.03*	0.084	2.37*	0.032
	t ₉	2.89*	0.024	3.71*	0.041	1.12	
manifeste Angst	t ₁	0.44		2.83*	0.014	0.24	
	t ₅	3.61*	0.031	1.72		1.13	
	t ₉	1.32		1.53		0.76	

* $p < .05$

Für die zu t_5 (erste Nachuntersuchung nach der Kommentierungsphase) erhobenen kognitiv-motivationalen Variablen ergaben sich in einer zweifaktoriellen, multivariaten Varianzanalyse (bifaktorielle MANOVA mit dem Pillai-Bartlett-V-Kriterium; siehe Bredenkamp, 1980; Bredenkamp & Erdfelder, 1984) sowohl für den Faktor „Leistungsstand zu t_1 “ ($F(6/352) = 12.47; p < .01$) als auch für den Faktor „Gruppenzugehörigkeit“ ($F(9/528) = 6.13; p < .01$) signifikante Haupteffekte, deren letzter zunächst für die Annahme der globalen Alternativhypothese (4) spricht: Die experimentellen Bedingungen führen zu bedeutsamen Unterschieden in den betrachteten situationsspezifischen kognitiv-motivationalen Variablen. Da aber auch der MANOVA-F-Wert für die Interaktion beider Faktoren signifikant ist ($F(33/504) = 3.44; p < .01$), muß dieser Haupteffekt differenziert interpretiert werden und Arbeitshypothese (5) wird global bestätigt. Dies zeigt sich differenzierter in den zweifaktoriellen, univariaten Varianzanalysen (ANOVAs), deren Ergebnisse in Tabelle 1 aufgeführt sind. Für die Daten von Erhebungszeitpunkt t_5 ergibt sich bei der Variablen „Valenz einer Verbesserung“ nur ein signifikanter Haupteffekt für den Leistungsstand des Schülers zu t_1 . Bei der Verbesserungs-Erwartung und der Verbesserungsmotivation zeigen sich dagegen auch statistisch bedeutsame Haupteffekte für die Gruppenzugehörigkeit und bedeutsame Interaktionseffekte. Während sich die Kommentierung somit auf die subjektive Wichtigkeit einer Verbesserung nicht differentiell auswirkt, werden solche Effekte für Verbesserungs-Erwartungen und -motivationen differenziert (für die verschiedenen Leistungsgruppen) bestätigt. Dies entspricht der Arbeitshypothese (5). Die Schätzwerte für die Effektstärke (w^2 -Werte in Tabelle 1) zeigen dabei, daß die Auswirkungen der experimentellen Bedingungen bei den Verbesserungs-Erwartungen etwas höher liegen als bei den Verbesserungsmotivationen, was dem konzeptuellen Gerüst erwartungswert-theoretischer Motivationstheorien entspricht, in denen ja multiplikative Beziehungen zwischen Valenz- und Erwartungsschätzungen für die Vorhersage von motivationalen Variablen postuliert werden. Nach den Kriterien Cohens (1977²) können die beobachteten Effektstärken für die Kommentierung — wie erwartet — als „klein“ bzw. „mittel“ beurteilt werden. In Abbildung 2 ist nun exemplarisch für die Variable „Verbesserungs-Erwartung“ (zu t_5) die Interaktion der Faktoren „Gruppenzugehörigkeit / Kommentierungsart“ und „Leistungsstand zu t_1 “ graphisch dargestellt. Die Abbildung verdeutlicht, daß Interaktionshypothese (5) inhaltlich voll bestätigt wird: Der sozial orientierte Kommentar führt bei leistungsschwächeren Schülern zu sehr geringen Verbesserungs-Erwartungen, bei leistungsstärkeren dagegen zu erhöhten; die Schere zwischen den Leistungsgruppen öffnet sich unter dieser Kommentierungsart also weiter. Die sachlich orientierte Lehrerkomentierung zeigt dagegen ebenso wie die individuell orientierte positive Effekte bei allen Leistungsgruppen, wobei der individuelle Kommentar vor allem bei den leistungsschwächsten Schülern sich signifikant stärker auf die Verbesserungs-Erwartungen auswirkt als der sachliche Lehrerkommentar.

Effekte auf Zeugnisnoten und deren subjektive Abbildungen. Als zweite Variablengruppe wurden die Zeugnisnoten und fachspezifische kognitiv-motivationale Variablen (Zufriedenheit mit der Note, Interesse an Mathematik, drei Einschätzungen des Selbstkonzepts eigener mathematischer Fähigkeiten; s. o.), die sich schon allgemeiner auf das Fach Mathematik beziehen als die situations-spezifischen, für die Auswertung zusammengefaßt. Bei der Voruntersuchung bestanden keine bedeutsamen Gruppenunterschiede in diesen Variablen, was in Tabelle 1 für einige von ihnen spezifisch dokumentiert ist.

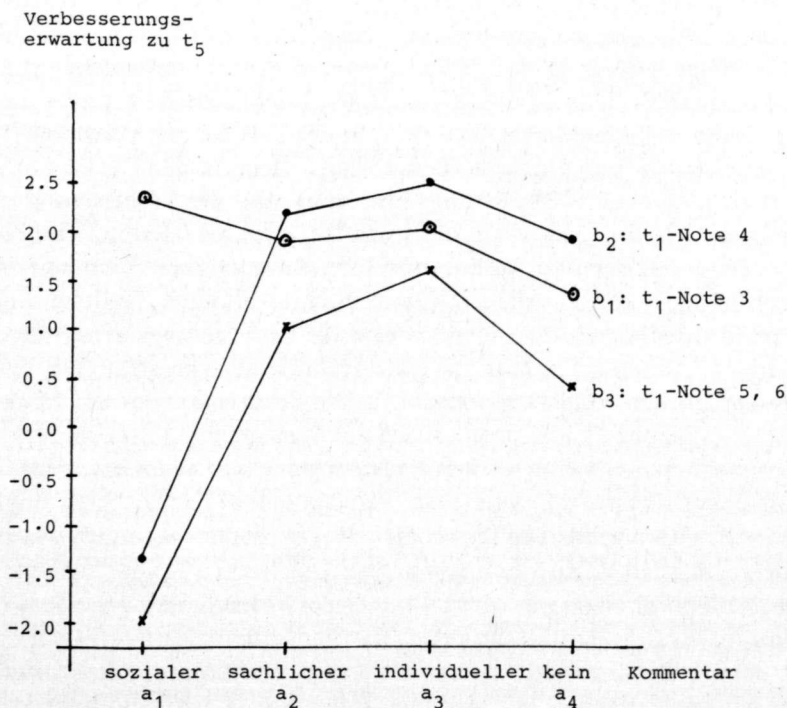


Abb. 2. Veranschaulichung der Interaktion „Kommentarart“ (A)
x „Leistungsstand zu t_1 “ (B)
für die Verbesserungserwartung in Mathematik zu t_5 .

Eine bifaktorielle MANOVA (mit den Faktoren „Leistungsstand zu t_1 “ und „Gruppenzugehörigkeit“) dieser sechs Variablen führte erneut zu signifikanten Haupteffekten ($F(18/747) = 19.37$ bzw. $F(18/747) = 14.91$; $p < .01$) und zu einem bedeutsamen Interaktionseffekt ($F(90/1440) = 4.37$; $p < .01$), was erneut Arbeitshypothese (5) global bestätigt: Auswirkungen verschiedener Art der Kommentierung zeigen sich je nach Leistungsstand der Schüler differentiell.

In dieser und den folgenden Analysen wurde stets mit vier Leistungsgruppen (bezogen auf t_1) gerechnet (Schüler mit „sehr gut/gut“, „befriedigend“, „ausreichend“ und „mangelhaft/ungenügend“ auf dem letzten Zeugnis), da sich weder bei Ausschluß der leistungsstarken noch bei Ausschluß der leistungsschwachen Schüler (auch erfaßt über die Frage nach der erlebten Überforderung im Mathematikunterricht) andere Ergebnisse ergaben.

Für drei der sechs analysierten fachspezifischen Variablen sind in Tabelle 1 die Ergebnisse der zweifaktoriellen ANOVAs für die Daten aus der ersten (t_5) und der zweiten Nachuntersuchung (t_9) wiedergegeben. Signifikante Haupt- und Interaktionseffekte belegen die Wirkung der Kommentierung für die Daten der ersten Nachuntersuchung. Exemplarisch für die Note auf dem Zwischenzeugnis (t_5) ist in Abbildung 3 die Interaktion der Faktoren „Gruppenzugehörigkeit / Kommentierungsart“ und „Leistungsstand (t_1)“ dargestellt. Inhaltlich ergeben sich sehr ähnliche Befunde wie bei den situationsspezifischen kognitiv-motivationalen Variablen. Während sich bei der sozial orientierten Kommentierung die Schere zwischen den Leistungsgruppen weiter öffnet, verbessern sich bei der sachlich und bei der individuell orientierten Kommentierung alle Schüler, wobei erneut die individuelle Kommentierungsart der sachlichen (vor allem bei leistungsschwachen Schülern; $p < .05$) überlegen ist. Die Schätzungen für die Effektstärke bewegen sich erneut im Bereich „kleiner“ bzw. „mittlerer“ Werte (nach Cohen, 1977²), wobei nur die des Faktors „Leistungs-

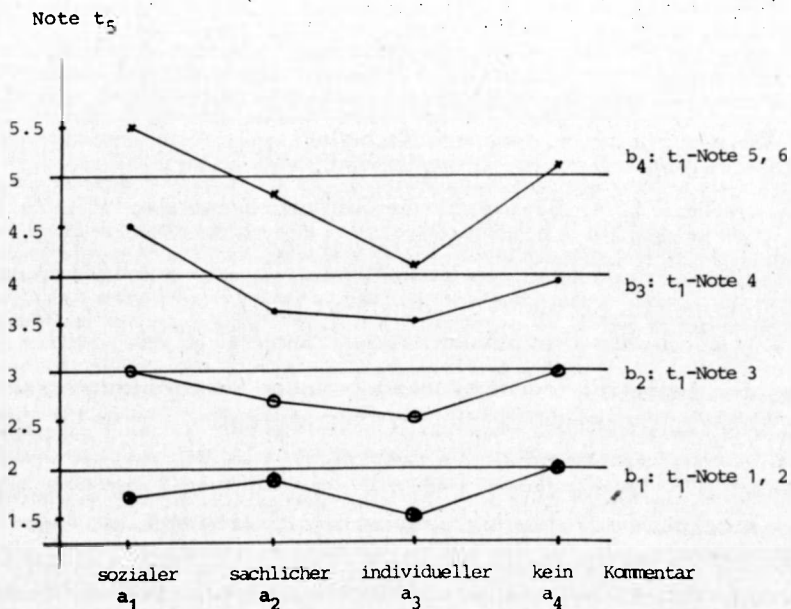


Abb. 3. Veranschaulichung der Interaktion „Kommentierungsart“ (A)
x „Leistungsstand zu t_1 “ (B)
für die Mathematikzensur auf dem Halbjahreszeugnis (t_5).

stand (t_1)“ herausfallen. Ihre Höhe entspricht weitgehend korrelativen Untersuchungsbefunden, die darauf deuten, daß die „besten“ Prädiktoren für Schulnoten frühere Schulnoten sind (siehe etwa Bierhoff-Alfermann, 1976; Schwarzer, 1979; Schneider, 1980).

Bei der zweiten Nacherhebung (t_9), die ein halbes Jahr (in dem nicht mehr kommentiert wurde) nach t_5 stattfand, sind alle Kommentierungseffekte wieder verschwunden (siehe Tabelle 1). Bestehen bleiben alleine die (abgeschwächten) Haupteffekte für den Faktor „Leistungsstand (t_1)“. Dies dokumentiert sich auch in den Ergebnissen einer bifaktoriellen MANOVA für die zu t_9 erhobenen Daten (Haupteffekt „Gruppenzugehörigkeit“: $F(18/747) = 0.53$; Haupteffekt „Leistungsstand“: $F(18/747) = 3.88, p < .01$; Interaktionseffekt: $F(90/1440) = 0.19$).

Interaktionshypothese (5) wird also auch durch die Befunde zu den fachspezifischen kognitiv-motivationalen Variablen und zu den Zeugnisnoten in Mathematik bestätigt; gleichzeitig wird aber festgestellt, daß sie nur (kurzfristig) für den Erhebungszeitpunkt t_5 unmittelbar nach der Kommentierungsphase gilt. Arbeitshypothese (3), die sich auf Nachwirkungen der Kommentierung (nach ihrer Absetzung; t_9) bezieht, muß dagegen für die fachspezifischen Variablen zurückgewiesen werden.

Effekte auf schulleistungsspezifische Persönlichkeits- und Einstellungsvariablen. Die drei Skalen des IPC-PL-Fragebogens zu Kontrollüberzeugungen im Bereich des Problemlösehandelns und die drei inhaltlichen Skalen des AFS, in denen zu t_1 keine bedeutsamen Gruppenunterschiede ermittelt werden konnten (s. o. und Tabelle 1), gingen ebenfalls als abhängige Variablen in bifaktorielle MANOVAs ein. Für die bei der ersten Nacherhebung (t_5) erhaltenen Daten ergaben sich wiederum signifikante Haupteffekte für den Leistungsstand (t_1 ; $F(18/747) = 16.27; p < .01$) und für die Gruppenzugehörigkeit ($F(18/747) = 10.38; p < .01$) sowie ein signifikanter Interaktionseffekt ($F(90/1440) = 4.21; p < .01$). Für die bei der zweiten Nacherhebung (t_9) erhobenen Daten ergaben sich signifikante (abgeschwächte) Haupteffekte ($F(18/747) = 2.77$ bzw. $1.85; p < .05$), jedoch kein bedeutsamer Interaktionseffekt ($F(90/1440) = 0.45$).

In Tabelle 1 sind die Ergebnisse der a posteriori durchgeführten zweifaktoriellen ANOVAs für beide Erhebungszeitpunkte getrennt dargestellt. Während sich für die Daten aus der ersten Nachuntersuchung für die Skalen zur Internalität, zur sozial bedingten Externalität, zur Schulunlust und zur Prüfungsangst erneut die bekannten Haupt- und Interaktionseffekte zeigen, die Arbeitshypothese (5) bestätigen, finden sich bei den Skalen zur fatalistisch bedingten Externalität und zur manifesten Angst nur bedeutsame Haupteffekte für die Gruppenzugehörigkeit. Mit Bezug auf Arbeitshypothese (3) kann ferner festgestellt werden, daß bei vier der sechs bereichsspezifischen Variablen (Internalität, soziale Externalität, Schulunlust und Prüfungsangst) Nachwirkungen der Kommen-

tierung beobachtet werden können, was sich an den signifikanten Haupteffekten für den Faktor „Gruppenzugehörigkeit“ für die t_9 -Daten zeigt.

Vergleiche der Effektstärken. In Arbeitshypothese (2) wird davon ausgegangen, daß die stärksten Effekte der Kommentierung bei den situationsspezifischen kognitiv-motivationalen Variablen, schwächere bei den fachspezifischen und noch schwächere bei den bereichsspezifischen Variablen zu beobachten sind. Der Vergleich der in Tabelle 1 aufgeführten w^2 -Werte, die als Schätzungen der Effektstärke dienen, zeigt, daß diese Hypothese akzeptiert werden kann: Sowohl für die Stärke der Haupteffekte des Faktors „Gruppenzugehörigkeit“ als auch für die der Interaktionseffekte ergaben sich für die Sequentierung der Variablen in Vorzeichentests (*one-sample runs test*) signifikante Abweichungen ($p < .05$; einseitige Testung) von einer Zufallsfolge der abhängigen Variablen. Festzuhalten bleibt aber, daß alle Effekte im „kleinen“ bzw. „mittleren“ Bereich angesiedelt sind ($w^2 \leq .12$). Bei Beachtung der Art und des Umfangs der experimentellen Manipulation ist dies im Vergleich zu umfassenderen Motivförderungsprogrammen durchaus im erwarteten Rahmen.

Effekte auf Noten in einzelnen Mathematikarbeiten. Nicht angesprochen wurden bislang die bei den Zwischentestungen (also nach der Rückgabe der Mathematikarbeiten) erhobenen Variablen. Exemplarisch seien hier die Ergebnisse für die Noten in den Klassenarbeiten kurz dargestellt. Es zeigte sich, daß sich die Noten zu t_2 zwischen den vier Gruppen nicht bedeutsam unterscheiden ($F(3/379) = 1.03$). Bei dieser Prüfungsarbeit handelt es sich ja auch um die erste, die in den Experimentalgruppen kommentiert zurückgegeben wurde. Schon zu t_3 (also bei der zweiten Arbeit in der Kommentierungsphase) zeigen sich aber bedeutsame Gruppenunterschiede ($F(3/378) = 7.19, p < .05; w^2 = .05$). Wie auch t-Tests für abhängige Stichproben belegen, führt die sachliche Kommentierung zu einer Verschlechterung der Durchschnittsnote, die individuelle dagegen zu einer Verbesserung. Zu t_4 kann eine weitere Verbesserung der Gruppe III (individueller Kommentar) und zudem eine erhebliche Verbesserung der Gruppe II (sachlicher Kommentar) beobachtet werden ($F(3/379) = 4.77, p < .05; w^2 = .03$). Diese Veränderungen und auch die Unterschiede zu den anderen Gruppen bleiben bei t_5 (Zeugnisnote; s. o.) und t_6 (erste Arbeit nach dem Absetzen der Kommentierung; $F(3/375) = 3.41, p < .05; w^2 = .02$) erhalten, bilden sich dann aber zu t_7 und t_8 zurück ($F(3/378) \leq 1.42$). Dies stimmt mit den Befunden zu den Zeugnisnoten (bezogen auf t_5 und t_9) überein. Interessant ist dabei, daß sich für Gruppe I (sozial orientierter Kommentar) keine statistisch bedeutsamen Veränderungen in den Noten zeigen. Dies kann durch die oben dargestellten Ergebnisse zur Interaktion des Leistungsstandes, der hier (aus darstellungstechnischen Gründen) nicht beachtet wird, und der Kommentierungsart erklärt werden: Die Verschlechterung der leistungsschwächeren Schüler und die weitere Verbesserung der leistungsstarken unter dieser Kommentierungsart heben sich in den Gesamtmittelwerten auf. Die zeitlich kumulative Wirkung von Lehrer-

kommentaren zu Noten, die in Hypothese (1) angenommen wird, kann somit durch diese Befunde im wesentlichen als gestützt betrachtet werden. Differentiell zeigt sich jedoch, daß eine „echte“ Kumulation der Effekte nur bei individuell orientierten Kommentaren festzustellen ist. Bei sachlichen Kommentaren kommt es dagegen zunächst zu einem Leistungseinbruch, der dann jedoch relativ rasch abgefangen wird. Nachwirkungen der Kommentierung (Arbeitshypothese 3) sind bei den Noten in Prüfungsarbeiten nicht feststellbar. Die Bedeutung der Interaktion von Leistungsstand des Schülers und Kommentarart für Effekte zeigt sich auch hier ganz deutlich bei den Schülern, die den sozial orientierten Kommentar erhalten haben.

Ergebnisanalysen auf Klassenebene. Die für die Arbeitshypothesen (4) und (5) erhaltenen Befunde aus Berechnungen, in denen das Individuum als Analyseeinheit diente, werden wegen der möglichen Kritik aufgrund des „groberen“ Randomisierungsprozesses (s. o.) durch Berechnungen abgesichert, in denen die Klassenmittelwerte als Analyseeinheit dienen. In eine zweifaktorielle Varianzanalyse für die Zeugnisnote zu t_5 gingen als Analyseeinheiten nicht die Einzelnoten der untersuchten Schüler sondern die Mittelwerte aus den 13 untersuchten Klassen ein. Als unabhängige Variablen dienten erneut der Leistungsstand zu t_1 und die Gruppenzugehörigkeit (also Kommentierungsart). Ebenso wie auf Individualebene (siehe Tabelle 1) ergeben sich auch hier bedeutsame Haupteffekte für die Kommentierungsart ($F(3/36) = 66.36, p < .05$) und den Leistungsstand ($F(3/36) = 1047.72, p < .05$) sowie ein signifikanter Interaktionseffekt ($F(9/36) = 10.00, p < .05$), der Arbeitshypothese (5) bestätigt. Auch die Schätzungen der Effektstärke für die Kommentierungsart ($w^2 = .056$) und für die Interaktion ($w^2 = .023$) sind in ähnlichen numerischen Bereichen angesiedelt. Nur der Wert ($w^2 = 0.905$) für den Haupteffekt „Leistungsstand“ steigt erheblich an, was darauf verweist, daß aufgrund klassenspezifischer Schülergruppierungen nach dem (früheren) Leistungsstand der aktuelle Leistungsstand dieser Schülergruppe mit sehr hoher Wahrscheinlichkeit richtig vorhergesagt werden kann. Ähnliche Phänomene, von denen man sich nicht verwirren lassen sollte, finden sich häufiger beim Wechsel der Analyseeinheit (siehe etwa auch Helmke, 1983), durch den eine andere Art der Information ausgenutzt wird.

Wegen der Verringerung der Beobachtungseinheiten N bei Analysen auf Klassenebene wurden ergänzend eine Kruskal-Wallis Varianzanalyse (H-Test) und Nemenyis D_{ij} -Tests für Einzelvergleiche berechnet. Durch diese nonparametrischen Verfahren konnte Arbeitshypothese (4) insofern bestätigt werden, als auf Klassenebene signifikante Gruppenunterschiede allgemein ($H_{corr}(df = 3) = 11.835; p < .01$) und spezifische Unterschiede in den Mathematiknoten auf dem Zwischenzeugnis zwischen den Gruppen III und I sowie III und IV ($D_{ij} \geq 19; p < .05$; einseitige Fragestellung) nachgewiesen werden konnten. Obwohl im bifaktoriellen H-Test nach Bredekamp (1974; siehe auch Lienert, 1978², S. 1036ff.) eine signifikante Interaktion zwischen Leistungsstand und Kommentarart nachgewiesen werden konnte, die der in den ANOVAs erzielten entspricht, soll auf deren Darstellung verzichtet werden, da neuerdings Erdfelder & Bredekamp (1984) darauf aufmerksam gemacht haben, daß dieses Verfahren nur unter äußerst restriktiven Annahmen (es dürfen keine bedeutsamen Haupteffekte vorliegen!) zu sinnvoll interpretierbaren Interaktionen führt. Zudem raten die Autoren allgemein von der Transformation erhobener Originaldaten auf Ordinalniveau ab.

Diskussion

Die vorgelegten Befunde zeigen, daß die schriftliche Kommentierung von Zensuren in Prüfungsarbeiten durch die Lehrer doch nicht eine so einfache Möglichkeit der Optimierung des Lern- und Leistungsverhaltens von Schülern ist, wie es von einigen Autoren (vor allem Page, 1958), deren Arbeiten häufig in Lehr- und Einführungsbüchern dargestellt werden, nahegelegt wird. Dafür spricht nicht nur die widersprüchliche Befundlage zur Wirkung von Lehrerkomentaren (siehe Stewart & White, 1976), sondern vor allem die Tatsache, daß in den bisherigen Arbeiten versäumt wurde, Wechselwirkungen zwischen dem Leistungsstand der Schüler und der Kommentarart zu analysieren. Stets wurde nur die globale Effektivität von Kommentaren bei allen an der Untersuchung beteiligten Schüler betrachtet. Die vorliegende Arbeit weist nun eindeutig darauf hin, daß solche allgemeinen (Haupt-)effekte für verschiedene Typen der Lehrerkomentierung zwar nachweisbar, wegen der empirisch bestätigten Interaktion mit dem Leistungsstand der Schüler aber differenziert interpretiert werden müssen. Im Sinne einer "aptitude-treatment interaction" ergab sich für nahezu alle betrachteten abhängigen Variablen, daß die Kommentarwirkungen vom Leistungsstand der Schüler abhängen. Damit kann die widersprüchliche Befundlage zu dieser Fragestellung aufgelöst werden: Nicht alleine der Inhalt des Kommentars ist für die Art seiner Wirkung ausschlaggebend (wie es Stewart & White, 1976, in ihren Schlußfolgerungen vermuten; hier: Arbeitshypothese 4), sondern vielmehr seine Wechselwirkung mit dem Leistungsniveau des Schülers, der die differenzierte Rückmeldung erhält (Arbeitshypothese 5). Die vorgelegten Befunde weisen klar darauf, daß in Ergänzung zu und Differenzierung von Arbeitshypothese (4) Arbeitshypothese (5) anzunehmen ist.

Die beobachteten differentiellen Effekte der Lehrerkomentierung können folgendermaßen zusammengefaßt werden: (1) Sozial orientierte Lehrerkomentare wirken bei leistungsschwächeren Schülern deutlich negativ, bei leistungstärkeren neutral oder leicht positiv; bei dieser Kommentierungsart öffnet sich somit die Schere zwischen den Schülern (mit Bezug zu einer Reihe von Variablen; s. u.) weiter. (2) An einem sachlichen (lehrstoff-bezogenen) Gütemaßstab orientierte Kommentare wirken in der Tendenz bei allen Schülern positiv, ohne daß gesagt werden kann, daß eine Leistungsgruppe von ihnen besonders profitiert; die Effekte sind jedoch eher gering. (3) Individuell orientierte Kommentare wirken ebenfalls bei allen Schülern tendenziell positiv, wobei aber insbesondere die leistungsschwachen am meisten von ihnen profitieren. In der vorliegenden Untersuchung hat zwar die über die Zensuren und die über die subjektiv von den Schülern erlebte Überforderung durch den Unterricht erfaßte potentielle Variabilität der abhängigen Variablen keine Rolle gespielt (was evtl. mit dem Schultyp und der Art der kognitiv-motivationalen Variablen erklärt werden kann). Es sei aber betont, daß solche differentiellen Kommentarwirkungen stets nur bei

gegebener potentieller Variabilität der Variablen erreicht werden können. Auch sei davor gewarnt, die individuell orientierte Kommentierung zu verabsolutieren (siehe etwa Henschel, 1890; Butterworth & Michael, 1975). Mehr oder weniger stereotype Wiederholungen können auch bei diesem Kommentartyp längerfristig zu negativen (bestenfalls neutralen) Folgen führen, was durch seine Konflikthaftigkeit zur in unserem Schulsystem dominierenden sozialen Bezugsnorm-orientierung verstärkt werden kann (siehe auch Rheinberg, 1980, S. 164 ff.). Es ist ja zu bedenken, daß (im öffentlichen Schulsystem) bei jedem Kommentartyp schon alleine durch die Vergabe der Noten soziale Vergleichsprozesse angesprochen und für die Schüler möglich werden. Durch die Kommentierung wird dies lediglich verstärkt (sozial orientierter Kommentar) oder abgeschwächt und differenziert (sachlich und individuell orientierter Kommentar).

Die Arbeitshypothesen (1) bis (3), die sich auf zeit- und variablenspezifische Wirkungen von Lehrerkomentaren beziehen, werden durch die vorgelegten Befunde weniger stringent gestützt als Hypothese (5). In den Analysen der zu den Zwischentestungen (nach der Rückgabe von Prüfungsarbeiten) erhobenen Variablen deuten sich zwar zeitlich-kumulative Kommentarwirkungen an, die aber nicht für alle drei Kommentartypen kontinuierlich beobachtet werden konnten (Arbeitshypothese 1). Zudem zeigt sich mit Bezug zur Arbeitshypothese 3, daß nach dem Absetzen der Kommentierung in den Experimentalgruppen (zweite Untersuchungsphase) die Effekte bei den Mathematikzensuren und den situations- und fachspezifischen kognitiv-motivationalen Variablen recht schnell wieder verschwinden. Lediglich bei den bereichsspezifischen Persönlichkeits- und Einstellungsvariablen (Kontrollüberzeugungen für Problemlösehandeln, Schulunlust, Prüfungsangst) können Nachwirkungen (zu t_9 , also ein halbes Jahr nach Absetzen der Kommentierung) registriert werden (Bestätigung von Arbeitshypothese 3). Vorher (zu t_3) bestehende Interaktionseffekte werden aber nicht mehr beobachtet. Inhaltlich bedeutet dies, daß die Zunahme der Schulunlust und der sozial bedingten Externalität in Gruppe I (sozialer Kommentar) auch langfristig erhalten bleibt. Analoges gilt für die Abnahme von fatalistisch bedingter Externalität und Prüfungsangst sowie die Zunahme der Internalität in Kontrollüberzeugungen in der Gruppe, deren Noten unter individueller Perspektive kommentiert worden ist. Allgemein kann dabei festgestellt werden, daß — bei bestehenden statistischen Signifikanzen — die Effektstärken eher als „klein“ bzw. „mittel“ (nach den Kriterien Cohens, 1977²) zu bewerten sind. Bedenkt man die Komplexität des Bedingungsgefüges der untersuchten abhängigen Variablen (siehe etwa Krapp & Mandel, 1976; Weinert & Petermann, 1980; Krug, 1983), so bleibt es aber doch wesentlich, daß durch eine isolierte Intervention, die sich auf einen einzigen situativen Faktor bezieht (Art der Leistungsrückmeldung in drei Prüfungsarbeiten), Effekte erreicht werden, die bis zu Effektstärken von 30 % erklärter Varianz reichen. Zu den variablenspezifischen Auswirkungen ist festzuhalten, daß sich — wie in Arbeitshypothese 5 erwartet — tendenziell die

deutlichsten Effekte bei den situationsspezifischen kognitiv-motivationalen Variablen (und da vor allem bei den subjektiven Verbesserungserwartungen) zeigen. Die Auswirkungen der Kommentierung auf die fachspezifischen und bereichsspezifischen Variablen sind etwas schwächer ausgeprägt. Dies bestätigt erneut (vgl. auch Krampen, 1979; Heckhausen & Rheinberg, 1980; Grobe & Hofer, 1983) den Nutzen der kognitivistischen Motivationstheorie und ihre konzeptuelle Erweiterbarkeit um ausgewählte dispositionelle Variablen (siehe auch Pekrun, 1983; Krampen, 1984a). Wichtig ist, daß gerade bei diesen generalisierten Variablen (Kontrollüberzeugungen, Prüfungsangst, Schulunlust) Nachwirkungen der Kommentierung festgestellt wurden. Ohne diese Befunde überbewerten zu wollen, sei darauf verwiesen, daß sich hier entwicklungspsychologische Implikationen auf tun. Die Untersuchung kann so auch als ein Beitrag zu der von Montada & Schmitt (1982) geforderten prospektiv und multivariat angelegten entwicklungspsychologischen Forschung bezeichnet werden, der — als Interventionsstudie — Defizite im Bedingungs- und Veränderungswissen zur Entwicklung von Kontrollüberzeugungen (siehe auch Matheny & Edwards, 1974; Krampen, 1982a, S. 135 ff.) und von Prüfungsangst (siehe auch Helmke, 1983; Pekrun, 1983) auffüllen hilft.

Abschließend sollen kurz Fragen nach der internen und externen Validität der Untersuchungsbefunde angesprochen werden. Durch das gewählte feldexperimentelle Vorgehen und das realisierte *Design* können die meisten der von Campbell & Stanley (1963; siehe auch Krauth, 1983) angeführten, die interne Validität gefährdenden Faktoren ausgeschlossen werden. Reaktive Versuchsbedingungen und somit Erwartungseffekte auf Schülerseite wurden ebenso wie bei Page (1958) und in den meisten Replikationsstudien (siehe im Überblick Stewart & White, 1976) dadurch vermieden, daß die Schüler nicht wußten, daß sie an einem Experiment teilnahmen, und sich die Kommentierung auf die in den normalen Unterrichtsablauf integrierten Mathematikarbeiten bezogen. Die von Rheinberg et al. (1979) festgestellte Abweichung des Unterrichts unter individueller Bezugsnormorientierung vom Gewöhnlichen wird hier eine geringere Rolle gespielt haben, da sich die experimentelle Manipulation nicht auf das gesamte Unterrichtsgeschehen (wie bei Rheinberg et al., 1979) bezog. Überdies ergaben Gespräche mit einzelnen Schülern keine Hinweise darauf, daß solche Abweichungen, die sich reaktiv ausgewirkt haben könnten, wahrgenommen wurden. Lehrererwartungen, die in den vorliegenden Arbeiten nur unzureichend beachtet wurden, können sich bei dem realisierten Vorgehen (jeder Lehrer verwendet nur eine Kommentarart bei allen Schülern einer Klasse) im Unterschied zu früheren Arbeiten (etwa von Page, 1958; Fittkau & Langer, 1974; Butterworth & Michael, 1975) nicht differentiell (d. h. für die Wirkungen der verschiedenen Kommentararten) ausgewirkt haben.

Problematisch bleibt alleine die Überprüfung der Nachwirkungen der Kommentierungsphase. Hier mögen Lehrererwartungen durchaus in dem Sinne eine

Rolle gespielt haben, daß nach Absetzen der Kommentierung jeweils für die Schüler der gesamten Klasse ein verändertes Unterrichtsverhalten gezeigt wurde. Durch den Einschluß einer Kontrollgruppe kann aber die Hypothese, daß Lehrerwartungen während der Kommentierungsphase differentiell gewirkt haben könnten, zurückgewiesen werden. Durch die Zufallsaufteilung der Klassen zu den experimentellen Bedingungen (mit den auswertungstechnischen Implikationen; s. o.) und die vor Beginn des Experiments gesicherte Vergleichbarkeit der Untersuchungsgruppen können auch alternative Erklärungen der berichteten Effekte weitgehend ausgeschlossen werden. Dies gilt insbesondere für die rivalisierenden Variablen der Bezugsnormorientierung der Lehrer, des Lern- und Leistungsklimas in den Klassen und alle erhobenen kognitiv-motivationalen Variablen.

Durch das feldexperimentelle Vorgehen ist gleichzeitig ein relativ hohes Maß an externer Validität gegeben, wobei freilich die Fragen nach der Generalisierbarkeit der Befunde für Schüler anderen Alters und anderer Schultypen offen bleiben. Zu bedenken ist auch, daß die hier realisierten „reinen“ Kommentare wohl kaum der schulischen Realität entsprechen. Häufiger werden Mischtypen sein, in denen jedoch aufgrund der vorgelegten Befunde inhaltliche Schwerpunktsetzungen (in Abhängigkeit vom Leistungsstand des Schülers) möglich werden. Es kann auch vermutet werden, daß eine partielle (also gelegentliche) Kommentierung der Zensuren u. U. effektiver ist als das hier empirisch überprüfte kontinuierliche Vorgehen (Kommentierung aller Arbeiten aller Schüler), da dadurch stereotype Wiederholungen etc. vermieden werden. Schriftliche Lehrerkommentare sollten also inhaltlich und zeitlich spezifisch eingesetzt werden, wobei der Leistungsstand des Schülers (eventuell auch sein subjektiv wahrgenommener Leistungsstand) beachtet werden muß. Nur dann — so ist aufgrund der vorgelegten Ergebnisse zu erwarten — werden Kommentare einen das Lern- und Leistungsverhalten fördernden Effekt haben. Die in der Pädagogischen Psychologie gängige Lehrmeinung, irgendwie geartete Lehrerkommentare hätten positive Effekte bei den Schülern, muß aufgegeben werden. Da (kommentierte) Rückmeldungen zu Handlungsergebnissen nicht nur in der Schule anzutreffen sind, sondern auch in der tertiären Sozialisation, in der Personalbeurteilung, in therapeutischen Interaktionen usw., ist zu vermuten, daß die vorgelegten Ergebnisse zur Interaktion von Rückmeldungs-/Kommentierungsart und Leistungsstand (erweitert: Fähigkeiten/Selbstkonzepte eigener Fähigkeiten etc.) auch diese Bereiche betreffende Implikationen aufweisen.

Summary

Presented is the analysis of the effects of three kinds of teacher comments (social-comparison, subject, and intraindividually oriented; control group) accompanied by marks in mathematics. Subjects were 385 students from grade 6 to 10. The assignment of an experimental condition was done randomly. During the first half of a year teachers

commented on the students' performance; in the second half, comments were not made in order to examine the duration of effects resulting from the comments made in the first half. Students were observed at the beginning of the experiment, at the end of the treatment period, and after the no-comment phase. The variables were grades, cognitive-motivational variables and some school-related personality characteristics. Analyses of variance revealed that there is an interaction between type of comment and the students' prior performance level for almost all dependent variables. The effects of comments with a duration longer than the experimental period were observed only for school-related personality variables (scope-specific locus of control, test anxiety). An outline of implications for teaching strategies as well as for developmental and educational psychology is given.

Literatur

- Bierhoff-Alfermann, D., Die Beziehung von Noten und Schülermerkmalen bei Schülern der 9. und 10. Klasse. *Psychologie in Erziehung und Unterricht*, 1976, **23**, 205—214.
- Bredenkamp, J., Nichtparametrische Prüfung von Wechselwirkungen. *Psychologische Beiträge*, 1974, **16**, 398—416.
- Bredenkamp, J., Theorie und Planung psychologischer Experimente. Darmstadt: Steinkopff, 1980.
- Bredenkamp, J. & Erdfelder, E., Multivariate Varianzanalysen nach dem V-Kriterium. *Psychologische Beiträge*, 1984 (zur Veröffentlichung eingereicht).
- Butterworth, T. W. & Michael, W. B., The relationship of reading achievement, school attitude, and self-responsibility behaviors of 6-grade pupils to comparative and individualized reporting systems. *Educational & Psychological Measurement*, 1975, **35**, 987—991.
- Campbell, D. T. & Stanley, J. C., Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Hg.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963 (Deutsche Bearbeitung in K. Ingenkamp (Hg.), *Strategien der Unterrichtsforschung*. Weinheim: Beltz, 1973, S. 99—194).
- Cohen, J., *Statistical power analysis for the behavioral sciences*. New York: Academic Press, 1977².
- Cook, T. D. & Campbell, D. T., *Quasi-experimentation. Design and analysis issues for field settings*. Chicago: Rand McNally, 1979.
- Erdfelder, E. & Bredenkamp, J., Kritik mehrfaktorieller Rangvarianzanalysen. *Psychologische Beiträge*, 1984 (im Druck).
- Fittkau, B. & Langer, H., Auswirkungen schriftlicher Ermutigungen unter Klassenarbeiten auf Angst und Leistungen der Schüler. *Psychologie in Erziehung und Unterricht*, 1974, **21**, 15—21.
- Gage, N. L. & Berliner, D. C., *Pädagogische Psychologie*. München: Urban & Schwarzenberg, 1977.
- Grobe, R. & Hofer, M., Kognitiv-motivationale Korrelate von Schulnoten: Typen motivierter Schüler. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1983, **15**, 292—316.
- Hays, W. L., *Statistics*. London: Holt, Rinehart & Winston, 1963.
- Heckhausen, H., *Motivation und Handeln*. Berlin: Springer, 1980.
- Heckhausen, H. & Rheinberg, F., Lernmotivation im Unterricht, erneut betrachtet. *Unterrichtswissenschaft*, 1980, **8**, 7—47.
- Helmke, A., *Schulische Leistungsangst*. Frankfurt/Main: P. Lang, 1983.
- Henschel, P., Ein Schülerbild. In W. Rein (Hg.), *Berichte aus dem Pädagogischen Universitätsseminar zu Jena*, Heft 2. Langensalza: Beyer & Söhne, 1890, S. 143—156.

- Matheny, K. B. & Edwards, C. R., Academic improvement through an experimental classroom management system. *Journal of School Psychology*, 1974, 12, 222—232.
- Mathis, B. C., Cotton, J. W. & Sechrest, L., *Psychological foundations of education*. New York: Academic Press, 1970.
- Kahl, T. N., *Unterrichtsforschung*. Kronberg/Taunus: Scriptor, 1977.
- Kahl, T. N., Buchmann, M. & Witte, E. H., Ein Fragebogen zur Schülerwahrnehmung unterrichtlicher Lernsituationen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1977, 9, 277—285.
- Klauer, K. J., Bezugsnormen zur Leistungsbewertung: Begriffe, Konzepte, Empfehlungen. In F. Rheinberg (Hg.), *Bezugsnormen zur Schulleistungsbewertung*. Jahrbuch für Empirische Erziehungswissenschaften. Düsseldorf: Schwann, 1982. S. 21—37.
- Krampen, G., Zur prognostischen Bedeutung kognitiv-motivationaler Effekte von Zensuren in einer Deutscharbeit bei Hauptschülern. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1979, 11, 250—260.
- Krampen, G., *Differentialpsychologie der Kontrollüberzeugungen*. Göttingen: Hogrefe, 1982a.
- Krampen, G., Zur Wirkung von Lehrerkommentaren bei der Bewertung schriftlicher Arbeiten. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1982b, 14, 58—64.
- Krampen, G., Handlungstheoretische Perspektiven zur Analyse politischer Partizipation und Apathie. In H. Moser & S. Preiser (Hg.), *Umweltprobleme und Arbeitslosigkeit*. Gesellschaftliche Herausforderungen an die Politische Psychologie. Weinheim: Beltz, 1984, S. 27—42. (a)
- Krampen, G., Feldexperimentelle Prüfung der Effekte von Lehrerkommentaren zu Zensuren in Prüfungsarbeiten auf Schüler: Untersuchungsansatz, Untersuchungsstand, Ergebnisse der Vorerhebungen. *Trierer Psychologische Berichte*, 1984, 11, Heft 2. (b)
- Krampen, G. & Lehmann, P., Zum Vorhersagewert der kognitiven Motivationspsychologie für schulische Zensuren, Verbesserungsmotivationen und Leistungsveränderungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1981, 13, 194—206.
- Krapp, A. & Mandl, H., Vorhersage und Erklärung der Schulleistungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1976, 8, 192—219.
- Krauth, J., Methodische Probleme in der pädagogischen Evaluationsforschung. *Zeitschrift für Empirische Pädagogik*, 1983, 7, 1—21.
- Krug, S., Motivförderungsprogramme: Möglichkeiten und Grenzen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1983, 15, 317—346.
- Lienert, G. A., *Verteilungsfreie Methoden in der Biostatistik*, Band II. Meisenheim/Glan: Hain, 1978².
- Lindquist, E. F., *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- Montada, L. & Schmitt, M., Systematik der Angewandten Entwicklungspsychologie: Probleme der Praxis, Beiträge der Forschung. In R. Oerter & L. Montada (Hg.), *Entwicklungspsychologie*. München: Urban & Schwarzenberg, 1982, S. 677—703.
- Page, E. B., Teacher comments and student performance. *Journal of Educational Psychology*, 1958, 49, 173—181.
- Pekrun, R., *Schulische Persönlichkeitsentwicklung*. Frankfurt/Main: Lang, 1983.
- Pekrun, R., An expectancy-value model of anxiety. In H. M. van der Ploeg, R. Schwarzer & C. D. Spielberger (Hg.), *Advances in test anxiety research*, Volume 3. Hillsdale, N. J.: Lawrence Erlbaum, 1984 (in press).

- Rheinberg, F., *Leistungsbewertung und Lernmotivation*. Göttingen: Hogrefe, 1980.
- Rheinberg, F., Kühmel, B. & Duschka, R., Experimentell variierte Schulleistungsbewertung und ihre motivationalen Folgen. *Zeitschrift für Empirische Pädagogik*, 1969, 3, 1—12.
- Schmidt, W., Darstellung und Verwendung der Leistungsmessung und Leistungsbeurteilung in der Schule. In E. Stephan & W. Schmidt (Hg.), *Messen und Beurteilen von Schülerleistungen*. München: Urban & Schwarzenberg, 1978, S. 153—171.
- Schneider, W., Kausalmodelle in der Pädagogischen Psychologie: Ergänzende Analysen in zwei einschlägigen Untersuchungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1980, 12, 88—95.
- Schwarzer, R., Sequentielle Prädiktion des Schulerfolges. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1979, 11, 170—180.
- Stewart, L. G. & White, M. A., Teacher comment, letter grades, and student performance. *Journal of Educational Psychology*, 1976, 68, 488—500.
- Tacke, G. & Lindner, F., Der Einfluß individualisierenden Lehrerverhaltens auf das Selbstkonzept von Schülern. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1981, 13, 190—193.
- Weinert, F. E. & Petermann, F., Erwartungswidrige Schulleistung oder unterschiedlich determinierte Schulleistung? In H. Heckhausen (Hg.), *Fähigkeit und Motivation in erwartungswidriger Schulleistung*. Göttingen: Hogrefe, 1980, S. 19—52.
- Westermann, R. & Hager, W., The relative importance of low significance level and high in multiple tests of significance. *Perceptual & Motor Skills*, 1983, 56, 407—413.
- Wieczerkowski, W., Nickel, H., Janowski, A., Fittkau, B. & Rauer, W., *Angstfragebogen für Schüler (AFS)*. Braunschweig: Westermann, 1975.
- Wottawa, H., Die Kunst der manipulativen Berichtlegung in der Evaluationsforschung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1981, 13, 45—60.
- Zielinski, W., Verfahren zur Beurteilung des Unterrichts. In F. E. Weinert, C. F. Graumann & M. Hofer (Hg.), *Funk-Kolleg Pädagogische Psychologie*, Band 2. Frankfurt/Main: Fischer, 1974, S. 901—923.

Anschrift des Verfassers:

Dipl.-Psych. Dr. Günter Krampen
 Universität Trier, FB I — Psychologie
 D — 5500 Trier