Klieme, Eckhard; Bürgermeister, Anika; Harks, Birgit; Blum, Werner; Leiß, Dominik; Rakoczy, Katrin

Leistungsbeurteilung und Kompetenzmodellierung im Mathematikunterricht. Projekt Co2CA

Klieme, Eckhard [Hrsg.]; Leutner, Detlev [Hrsg.]; Kenk, Martina [Hrsg.]: Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. Weinheim ; Basel : Beltz 2010, S. 64-74. - (Zeitschrift für Pädagogik, Beiheft; 56)

Zeitschrift für Pädagogik · 56. Beiheft

Kompetenzmodellierung

Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes

Herausgegeben von
Eckhard Klieme, Detlev Leutner und Martina Kenk

BELTZ

Herstellung: Lore Amann
Gesamtherstellung: Druckhaus „Thomas Müntzer“, Bad Langensalza
Printed in Germany
ISSN 0514-2717
Bestell-Nr. 41157

Inhaltsverzeichnis

Mathematische Kompetenzen

Naturwissenschaftliche Kompetenzen

Sprachliche und Lesekompetenzen

Fächerübergreifende Kompetenzen

Lehrerkompetenzen

Eckhard Klieme/Anika Bürgermeister/Birgit Harks/Werner Blum/Dominik Leiß/ Katrin Rakoczy

Leistungsbeurteilung und Kompetenzmodellierung im Mathematikunterricht

Projekt Co²CA[1]

Einleitung

Das Forschungsprojekt „Conditions and Consequences of Classroom Assessment (Co²CA)“[2] untersucht, in welchem Verhältnis Leistungsmessung, -bewertung und -beurteilung[3] zum Unterricht und zum Lernprozess der Schüler[4] stehen. Die alltägliche Praxis der Leistungsbeurteilung – von der informellen Fehlerdiagnose im Unterrichtsgespräch bis zu Kriterien und Verfahrensweisen der Notengebung – spiegelt einerseits die Ziele und Prozessqualitäten des Unterrichts, andererseits wirkt sie sich ihrerseits auf das Unterrichtsgeschehen sowie die kognitive und motivationale Entwicklung der Lernenden aus. In diese Praxis greifen neuerdings extern entwickelte Tests und Vergleichsarbeiten ein, die sich auf Bildungsstandards beziehen. In diesem Kontext untersucht das Projekt, wie verschiedene Formen des Assessments[5] genutzt werden und welche Wirkung sie entfalten. Es konzentriert sich dabei auf den Mathematikunterricht in mittleren Bildungsgängen der neunten Jahrgangsstufe.

1. Aktuelle Forschungen zur Leistungsbeurteilung im Unterricht

1.1 Formative Leistungsbeurteilung

In der angelsächsischen Fachsprache sind mit „formative assessment“ alle Formen von Leistungsbeurteilung gemeint, die Informationen über die Diskrepanz zwischen Lernzie-

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: KL 1057/10) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

2 2007–2009 gefördert unter KL 1057/10; seit Herbst 2009 gemeinsam geleitet von E. Klieme, K. Rakoczy, W. Blum und D. Leiß.

3 Wir betrachten hier „Leistungsbeurteilung“ (englisch: Assessment) als einen Vorgang, der sowohl die Feststellung von Leistungen – bei quantitativen Verfahren mit nachweisbarer Güte als Messung bezeichnet – als auch deren normative Bewertung umschließt.

4 Aus Gründen der Lesbarkeit wird in diesem Beitrag nur die männliche Form verwendet, auch wenn beide Geschlechter gemeint sind.

5 Wir bezeichnen sowohl lernprozessbegleitende (formative) als auch bilanzierende (summative) Leistungsbeurteilung als Assessment – in Übereinstimmung mit der aktuellen angloamerikanischen Literatur, aber anders als im Rahmenantrag des DFG-Schwerpunktprogramms von 2005, wo dieser Begriff auf summative Erhebungen eingeschränkt wurde.

len und aktuellem Lernstand liefern und dadurch den Lehrenden und/oder den Lernenden selbst helfen, den weiteren Lernprozess zu gestalten (vgl. Sadler 1989; Black/William 1998). Was dies genau bedeutet, wird durchaus kontrovers diskutiert. Während einige Autoren (vgl. z.B. Hattie/Timperley 2007) betonen, dass jedes Testverfahren – auch ein breit angelegter standardisierter Multiple Choice-Test – formativ genutzt werden kann, bringen andere Autoren das Konzept in enge Verbindung mit spezifischen Unterrichtssituationen: diskursiven Sequenzen und Aufgabenstellungen, mit denen Wissen und Verständnis von Schülern offen gelegt werden können (vgl. Heritage 2007; Shavelson u.a. 2008).

Heritage und Shavelson unterscheiden dementsprechend verschiedene Arten von formativem Assessment: spontane Sequenzen („on the fly assessment"), geplante Frage-Antwort-Sequenzen, die diagnostische Informationen liefern („planned for interaction"), sowie stärker formalisiertes „curriculum-embedded assessment". Die beiden erstgenannten Varianten stellen diagnostische Situationen dar, in denen Leistungen von Schülern evoziert, beobachtet, interpretiert und durch die Lehrkraft kommentiert werden. Hier wird also adaptiv unterrichtet; eine explizite Messung und Bewertung von Schülerleistungen findet jedoch nicht statt.[6] Im Rahmen des DFG-Schwerpunktprogramms interessiert daher eher die stärker formalisierte Art der Leistungsmessung und -bewertung, also das „curriculum-embedded assessment". Auch hierfür finden sich in der einschlägigen Literatur und der Testpraxis in den USA ganz unterschiedliche Beispiele: Kommerzielle Testanbieter bieten Varianten ihrer summativen, standardbezogenen Verfahren als „benchmark tests" an, die beispielsweise quartalsweise den Lernfortschritt festhalten sollen. Die Tests der Projekte des Berkeley Evaluation and Assessment Research (BEAR) Center (vgl. Wilson 2008) und „Power Source" (vgl. Baker 2007) sind hingegen sehr unterrichtsnah angelegt. Während Wilson komplexe offene Aufgaben stellt, aus deren Beantwortung auf die Entwicklung des Verständnisses geschlossen werden soll, arbeiten Baker und Mitarbeiter mit Serien von Tests, die jeweils spezifische Zielbereiche prüfen. Shavelson und Mitautoren (2008) haben zunächst mit ähnlichen formalisierten Tests gearbeitet, mussten aber erfahren, dass die Lehrkräfte aufwändige Testserien nicht in ihren Unterricht integrieren konnten. Sie vermeiden inzwischen den „Assessment"-Begriff und sprechen stattdessen von „reflective lessons".

Einigkeit besteht aber darin, dass Durchführung, Interpretation und weiterführende Nutzung unterrichtsbezogener (formativer) Leistungsbeurteilung spezifische Lehrerkompetenzen erfordert und daher nicht ohne ein Lehrertraining eingeführt werden kann, das neben diagnostischen und instruktionalen Techniken auch ein vertieftes Verständnis fachlicher Lerninhalte und -prozesse erfordert. Dabei müsste auch die Notengebung (vgl. Brookhart 1993) als – zumindest im deutschen Schulsystem – wichtigste Art der formativen Leistungsbeurteilung berücksichtigen werden, die vermutlich von Wissen und Einstellungen der Lehrkräfte (vgl. Rakoczy u.a. 2008) sowie deren diagnostischer Kompetenz (vgl. Schrader 2006) beeinflusst ist und bestimmte Schülergruppen leicht verzerrt bewertet (vgl. Klieme 2003).

---

6 Ob es sich hier überhaupt um Assessment handelt oder „nur" um diskursiven Unterricht, wäre im Einzelfall daran zu prüfen, o bzumindest eine implizite Bewertung von Leistungen stattfindet.

1.2 Abgrenzung zur summativen Leistungsbeurteilung

In Abgrenzung zum formativen Assessment orientieren sich summative Leistungsbeurteilungen, die einen Bildungsabschnitt bilanzieren sollen, in den USA wie in Deutschland heutzutage an (Bildungs-)Standards. Standards geben landesweit die Kompetenzbereiche (Dimensionen), die Messeinheiten (Skalen und Stufen) und die „Sollwerte" (Minimal- oder Regelstandards) an, nach denen Schüler beurteilt werden.

Summatives und formatives Assessment können sich durchaus auf dieselben Kompetenzmodelle stützen und sollten es sogar, um die oft beschworene Passung („Alignment") zwischen Unterrichtsprozessen und Standards zu erreichen (vgl. Pellegrino/Chudowsky/Glaser 2001). Dennoch bleibt ein Spannungsverhältnis bestehen. Zum einen sind die Standards und die darauf bezogenen Tests wesentlich breiter angelegt als unterrichtsbegleitende Messungen. Zum anderen signalisieren Standards, dass Lehrende wie auch Lernende für die Erreichung der betreffenden Sollwerte verantwortlich gemacht werden, was zur Einengung von Unterrichtsprozessen führen und die Motivation der Schüler beeinträchtigen kann (vgl. Deci u.a. 1981; Koretz 2008). Zudem haben Rückmeldungen aus summativen Tests – sofern sie überhaupt gegeben werden – häufig eine weniger unterstützende Form.

1.3 Feedback als zentrales Element von formativem Assessment

Schon Sadler (1989) sah die Information über festgestellte Leistungen, die Beteiligten rückgemeldet wird, als Kernelement des formativen Assessments. Hattie und Timperley (2007) kritisieren an gängigen Assessmentsystemen, dass diese nur Momentaufnahmen des Lernstands abbilden, anstatt Informationen bereit zu stellen, die von den Schülern für den weiteren Lernprozess und von Lehrkräften für die Unterrichtsgestaltung genutzt werden können. Um ein tieferes Verständnis des Lerngegenstands zu erreichen, sollte Feedback konkrete Aussagen darüber machen, wie man dem Lernziel noch näher kommen kann, indem Bearbeitungsprozesse nachvollzogen, Fehler und Lücken identifiziert und Strategien benannt werden. Wir sprechen hier im Folgenden von prozessbezogenem Feedback.

Im Rahmen der Cognitive Evaluation Theory (vgl. Deci/Koestner/Ryan 1999) wird zwischen „informierendem" und „kontrollierendem" Feedback unterschieden. Von informierendem Feedback wird eine positive Wirkung auf Motivation und Leistung erwartet, da es durch die Information über die individuellen Kompetenzen der Lernenden das grundlegende Bedürfnis nach Kompetenz unterstützt. Operationalisiert wird informierendes Feedback häufig durch Formulierungen, die die individuelle Leistung der Lernenden mit der durchschnittlichen Leistung in der jeweiligen Lerngruppe (soziale Bezugsnorm), dem vorherigen Leistungsstand (individuelle Bezugsnorm) oder dem Leistungsziel (kriteriale Bezugsnorm, z.B. angestrebte Kompetenzstufe) in Beziehung setzen. Kontrollierendes Feedback hingegen betont, wie sich Lernende (hätten) verhalten sollen. Es kann damit als Bedrohung des Bedürfnisses nach Autonomie wahrgenommen werden und die Motivations- und Leistungsentwicklung beeinträchtigen.

Aus der Forschungsliteratur lassen sich auch Hypothesen darüber ableiten, welche Faktoren die Verarbeitung des Feedbacks bestimmen. Hierzu zählen u.a. die Attribution des Erfolgs bzw. Misserfolgs und die Anstrengungsbereitschaft der Lernenden.

2. Zielsetzungen und Vorgehensweise des Forschungsprojekts Co²CA

2.1 Phasen des Projekts

In dem auf insgesamt sechs Jahre angelegten Projekt sind unterschiedliche empirische Zugänge zum Thema „Leistungsbeurteilung und Feedback im Mathematikunterricht" vorgesehen:

1. Die *sekundäranalytische Auswertung* einer *videobasierten Unterrichtsstudie* (vgl. Klieme/Pauli/Reusser 2009) im Hinblick auf verbales Feedback im Unterricht sowie Notengebung (vgl. Rakoczy u.a. 2008).
2. Eine breite Test- und Befragungsstudie (*Survey*), in der zum einen Tests und Kompetenzmodelle für spezifische mathematische Unterrichtseinheiten entwickelt werden und zum anderen die Praxis der Leistungsbeurteilung im Unterricht mittels Lehrer- und Schülerbefragungen untersucht wird.
3. Eine mit dem Survey verbundene experimentelle Vorstudie *(Rückmeldestudie)*, bei der Schülern individuelle schriftliche Rückmeldungen gegeben werden, deren Akzeptanz und Wirkung auf die Attribution von Erfolg bzw. Misserfolg geprüft wird.
4. Ein *„Labor"-Experiment,* bei dem untersucht wird, wie sich die Breite des Testinhalts („formativer" Test, bezogen auf eine Unterrichtseinheit vs. „summativer" Test, bezogen auf Bildungsstandards allgemein) und die Art der Rückmeldung (kriterial, sozial vergleichend oder prozessbezogen) auf Motivation und Leistung von Schülern auswirken.
5. Eine *Interventionsstudie (Feldexperiment)*, bei der Lehrkräfte gezielt für unterschiedliche Arten des formativen Assessments trainiert werden. Hier soll ökologisch valide überprüft werden, ob sich positive Effekte von Feedbackformen aus dem „Labor" in die Praxis übertragen lassen.

Im Folgenden gehen wir ausschließlich auf die Schritte 2 und 3 ein, die den Kern der ersten Projektphase bildeten.

2.2 Anlage des Surveys mit Rückmeldestudie

Im Mittelpunkt der Projektphase 2007–2009 stand eine Studie in 66 Realschulklassen und Gesamtschulkursen, die zum mittleren Abschluss führen. Sie verfolgte vor allem das Ziel, Testaufgaben zu entwickeln und psychometrisch zu skalieren sowie spezifische Kompetenzmodelle aufzustellen, die später im Laborexperiment und in der Inter-

ventionsstudie eingesetzt werden können. Bei der Kompetenzmodellierung sollte speziell untersucht werden, wie differenziert sich mathematische Teilkompetenzen (hier: Modellierungskompetenz und technische Kompetenz) innerhalb eingegrenzter Themenbereiche (hier: Satzgruppe des Pythagoras sowie Lineare Gleichungssysteme) erfassen lassen und wie sich diese Maße zu den breiten Kompetenzmodellen verhalten, die bei Bildungsstandards und Vergleichsarbeiten benutzt werden.

138 *Mathematikaufgaben* zu den beiden Themenbereichen und den zwei Kompetenzdimensionen wurden zum Teil unter Rückgriff auf Vorprojekte entwickelt; hinzu kamen 38 Items aus den Erhebungen zu nationalen Bildungsstandards[7]. Die Aufgaben wurden nach einem Youden-Square-Design auf 31 Testhefte verteilt, sodass beliebige Kombinationen vorkamen und die Position der Aufgaben rotiert wurde. Insgesamt 1560 Schüler der neunten Jahrgangsstufe bearbeiteten nach Zufall eines der Testhefte. Die doppelstündige Erhebung im jeweiligen Klassenverband erfolgte zwischen Mai und Juni 2008 unter der Verantwortung externer Testleiter. Offene Antworten wurden von trainierten Studierenden der Mathematikdidaktik ausgewertet. Zu allen Aufgaben wurden stichprobenartig Zweitkodierungen angefertigt; bei ungenügender Kodiererübereinstimmung wurden alle vorliegenden Schülerlösungen nachkodiert. Abschließend lag Kappa über alle Aufgaben und Kodierer hinweg bei sehr guten .93. Die Aufgaben wurden mit Hilfe der Software Conquest gemeinsam skaliert; einige zumeist sehr schwierige Items mussten aufgrund mangelnder Passung ausgeschlossen werden.

Im begleitenden *Schülerfragebogen* wurde vor dem Test die Testmotivation erhoben (Beispielitem: „Ich bin fest entschlossen, mich bei diesem Test voll anzustrengen."), danach wurden u.a. die wahrgenommene Bezugsnormorientierung der Lehrperson (Skalen „kriteriale Bezugsnormorientierung", „individuelle Bezugsnormorientierung") und Hintergrundvariablen wie Geschlecht und sozialer Status („Bücherfrage") erfasst. Im *Lehrerfragebogen* sollte für jeden Schüler das Ergebnis der Bearbeitung (richtig/falsch) bei vier Beispielaufgaben prognostiziert werden. Der Anteil korrekter Prognosen wurde als Indikator der diagnostischen Kompetenz verwendet. Ergänzend sollten die Lehrkräfte ihr eigenes professionelles Wissen im Bereich Diagnostik einschätzen (Markieritem: „Ich besuche Weiterbildungen oder informiere mich in der Literatur zum Thema Leistungsbeurteilung/Benotungskriterien."). Darüber hinaus wurden Aspekte der Praxis der Leistungsbeurteilung erhoben. Der Fragebogen wurde von 46 Lehrkräften vollständig ausgefüllt.

Etwa ein halbes Jahr nach dem Survey wurde eine *Rückmeldestudie* durchgeführt. Dabei wurde 167 Schülern aus 14 Klassen die individuelle Testleistung rückgemeldet, und zwar nach Zufall in einer von drei Formen: sozial vergleichend, kriterial oder prozessbezogen. Die *sozial vergleichende Bedingung* beinhaltete den Vergleich der individuellen Schülerleistung mit der durchschnittlichen Leistung der Klasse, getrennt für Modellierungs- und technische Kompetenz. In der *kriterialen Bedingung* wurde die Schülerleistung anhand von Kompetenzstufenmodellen – wiederum getrennt für Modellierungs- und technische Kompetenz – mit dem Lernziel für Realschüler der neunten

7 Wir danken dem Institut zur Qualitätsentwicklung im Bildungswesen (Humboldt-Universität zu Berlin) für die Überlassung dieses Materials.

Jahrgangsstufe verglichen. In der *prozessbezogenen Bedingung* wurden anhand von Beispielaufgaben spezifische Stärken und Schwächen sowie entsprechender Verbesserungs- und Übungsbedarf für beide Kompetenzdimensionen aufgezeigt. Anschließend wurde mittels Fragebogen die Wirkung des Feedbacks auf emotionale und motivationale Variablen (z.B. Zufriedenheit mit dem Ergebnis und Attribution) erhoben.

2.3 Fragestellungen

Mit Hilfe von Daten der Survey- und Rückmeldestudie sollen folgende Forschungsfragen beantwortet werden:

1. Bilden unterrichtsbezogene Tests auf der einen und allgemeine Bildungsstandard-bezogene Tests auf der anderen Seite psychometrisch eigenständige Leistungsdimensionen ab?
2. Welche Beurteilungspraxis dominiert im Alltag des Mathematikunterrichts? Wie stark unterscheiden sich dabei einzelne Klassen und inwieweit beeinflussen Lehrermerkmale wie z.B. diagnostische Kompetenz die Beurteilungspraxis?
3. Welche Formen der Leistungsbeurteilung hängen mit guten Leistungen bzw. hoher Motivation zusammen?
4. Wie wirken sich prozessbezogenes, kriteriales und sozial vergleichendes Feedback auf Zufriedenheit und Attribution aus?

3. Ergebnisse

3.1 Zur Modellierung mathematischer Kompetenzen für formatives und summatives Assessment

139 Testaufgaben aus der Survey-Studie konnten mit ausreichend gutem Modell-Fit auf einer gemeinsamen latenten Dimension abgebildet werden. Sie decken zwei Kompetenzen (technische bzw. Modellierungskompetenz) und drei Themenbereiche (Satzgruppe des Pythagoras, lineare Gleichungssysteme sowie allgemeine Bildungsstandard-bezogene Themen) ab. Dies spricht dafür, dass es prinzipiell möglich ist, themenspezifische (unterrichtsbezogene) und breit angelegte, standardbezogene Aufgaben in gemeinsamen Kompetenzmodellen abzubilden.

Auf der Basis von Aufgabenanalysen konnten aber auch spezifische Kompetenzmodelle für Teildimensionen (z.B. den Themenbereich „Pythagoras“) entwickelt werden. Die empirischen Daten ließen sich mit mehrdimensionalen Modellen besser abbilden als mit dem globalen eindimensionalen Modell. Die Leistungen in den beiden thematisch eingegrenzten Testteilen korrelierten messfehlerbereinigt untereinander zu .66, aber mit den Leistungen bei Bildungsstandard-Aufgaben zu .81 bzw. .77. Dies spricht dafür, dass thematisch fokussierte Leistungsmessungen, wie sie für formatives Assess-

ment gebraucht werden, zusätzliche und spezifische Informationen enthalten, die nicht mit einer summativen, standardbezogenen Messung abgedeckt sind.

3.2 Zur Praxis der Leistungsbeurteilung im Unterricht aus Lehrer- und Schülersicht

Aus Angaben der *Lehrkräfte* zur Praxis ihrer Leistungsbeurteilung lassen sich drei ausreichend reliable Skalen bilden, die untereinander nur geringfügig korrelieren:

- *Verbale Rückmeldungen* sind sehr häufig; sie werden beispielsweise bei der Tafelarbeit in zwei Drittel aller Klassen „immer" gegeben.
- Weniger häufig sind Maßnahmen, in denen es um die Vergabe von Noten oder zumindest um die explizite Bewertung einer Haus- oder Unterrichtsaufgabe durch die Lehrkraft geht (*lehrerzentrierte Beurteilungspraxis*).
- Verschiedene Praktiken, die eine *aktive Partizipation von Schülern* beinhalten (z.B. Selbsteinschätzung, Peer-Bewertung), bilden eine weitere Skala. Die Mittelwerte ihrer Items liegen jedoch zwischen „nie" und „manchmal". Beispielsweise findet Leistungsbewertung als expliziter Gegenstand des Unterrichts („Die Schüler bewerten ihre Arbeit anhand von Kriterien, die wir im Unterricht entwickelt haben.") nur in jeder fünften Mathematikklasse „manchmal" statt; noch seltener werden Portfolios oder Lerntagebücher eingesetzt.

Etwa die Hälfte der Befragten gibt an, Schüler mit Migrationshintergrund zumindest manchmal besonders mild zu bewerten. Die in manchen empirischen Studien (vgl. z.B. Klieme 2003) gefundene „positive Diskriminierung" ist also keineswegs eine implizite, sondern vielfach auch eine explizite Strategie – und zwar geschlechtsabhängig: Lehrerinnen neigen stärker als Lehrer dazu, Migranten oder auch Mädchen milder zu beurteilen. Darüber hinaus tendieren sie eher zu partizipativen Beurteilungspraktiken und verbalen Rückmeldungen.

Die befragten *Schüler* nehmen zumeist eine individuelle Bezugsnormorientierung der Lehrkräfte wahr („Wenn jemand seine Leistungen gegenüber früher verbessert, so wird er dafür von unserem Lehrer besonders gelobt."). Je stärker die Lehrkraft selbst die Rückmeldefunktion von Noten betont, und je stärker sie von partizipativen Formen der Leistungsbeurteilung berichtet, umso eher berichten die Schüler von dieser Art der Bezugsnormorientierung. Nur etwa 10% der Lehrkräfte definieren jedoch vor Klassenarbeiten bzw. Prüfungen explizite Kriterien.

Die *diagnostische Kompetenz* der Lehrkräfte (d.h. hier: die Treffsicherheit ihrer Prognose von Testleistungen) erweist sich als durchaus verhaltensrelevant: sie korrespondiert mit der Tendenz, Schülern verstärkt verbale Rückmeldungen zu geben. Selbst eingeschätztes diagnostisches Wissen hängt mit der „objektiven" diagnostischen Kompetenz nicht zusammen, geht aber mit partizipativen Beurteilungsverfahren und (aus Sicht der Schüler) mit einer individuellen Bezugsnormorientierung einher.

3.3 Zusammenhänge der Beurteilungspraxis mit Motivation und Leistung der Schüler

Nachdem somit gezeigt ist, dass unterschiedliche Praktiken der Leistungsbeurteilung identifiziert werden können, die mit Lehrermerkmalen korrelieren, soll die für das Forschungsprojekt zentrale Frage untersucht werden, wie diese Variablen ihrerseits mit Motivation und Leistung der Schüler zusammenhängen. Wir prüften dies mit Hilfe von Mehrebenenanalysen (s. Tabelle 1) und gingen davon aus, dass die diagnostische Kompetenz sowie das mathematische Anspruchs- und somit Anregungsniveau des Unterrichts[8] hinsichtlich der motivationalen Prozesse neutral sind, aber eine bessere kognitive Förderung ermöglichen. Allerdings liegen hier ausschließlich querschnittliche Daten vor, sodass keine Kausalzuschreibungen zulässig sind.

Einflussgröße	**Effekt (*β*-Koeffizient) auf …**	
	Testmotivation	**Testleistung**
Ebene 2: Klasse		
Diagnostische Kompetenz der Lehrperson	−0,01	0,02 *
Lehrerzentrierte Beurteilungspraxis	−0,16 *	−0,33 *
Anspruchsniveau des Unterrichts	0,01	0,26 *
Ebene 1: Schüler		
Geschlecht männlich	−0,25 ***	0,36 ***
Sozialer Status	0,01	0,11 ***
Familiensprache Deutsch	0,02	−0,29 ***
Schülerperzeption: individuelle Bezugsnorm	0,17 ***	0,13 **

* Angegeben sind standardisierte Regressionskoeffizienten.

*Tab. 1: Mehrebenenanalysen zu Effekten der Leistungsbeurteilungspraxis**

8 Wir verwenden hierfür einen Indikator, der auf Schülerwahrnehmungen aufbaut: den Klassenmittelwert bezüglich der Frage „Wie häufig löst ihr im Mathematikunterricht Textaufgaben?“. Da Problemlöse- und Modellierungsaufgaben im deutschen Mathematikunterricht umgangssprachlich als „Textaufgaben“ bezeichnet werden, spiegelt sich in diesem Indikator das von Schülern wahrgenommene Anspruchsniveau des Unterrichts.

Leistungsbeurteilung und Motivation: Eine stark lehrer- und notenzentrierte Leistungsbeurteilung ist in der Tat mit niedrigerer Testmotivation verbunden, während der persönliche Eindruck von Schülern, die Lehrperson orientiere sich an individuellen Lernfortschritten, mit höherer Motivation einhergeht.

Leistungsbeurteilung und Testleistung: Als Leistungskriterium wurde der Mittelwert aus der Leistung in den Bereichen Modellierungskompetenz und technische Kompetenz verwendet. Auch hier zeigt sich, dass lehrerzentrierte Leistungsbeurteilung negativ, die Wahrnehmung einer individuellen Bezugsnorm jedoch positiv mit der Testleistung zusammenhängt, und die Effekte des individuellen Hintergrunds bei der Leistung – wie aus der Unterrichtsforschung hinlänglich bekannt – noch stärker ausgeprägt sind als bei der Motivation. Anders als bei der Erklärung von Testmotivation sind nun aber auch diagnostische Kompetenz und mathematisches Anspruchsniveau wichtig.

3.4 Auswirkungen von Feedback

Auf der Grundlage des Forschungsstands zu Feedback wurde erwartet, dass prozessorientierte und kriteriale Rückmeldungen in motivationaler Hinsicht positiver aufgenommen werden als eine sozial vergleichende Rückmeldung. Außerdem wurde vermutet, dass die Art der Rückmeldung sich darauf auswirkt, ob Erfolg bzw. Misserfolg eher auf Begabung oder Zufall zurückgeführt, also intern oder extern attribuiert werden.

In der Tat ergaben die Analysen, dass eine *kriteriale* Form der Rückmeldung, verglichen mit sozial vergleichendem Feedback, die Zufriedenheit mit dem Rückmeldeergebnis und die Tendenz zur internalen Attribution („Begabung“ statt „Glück“) verstärkt. Allerdings zeigen sich derartige Effekte nicht bei der *prozessbezogenen* Rückmeldung. Dies könnte vor allem damit zusammenhängen, dass zwischen Test und Feedback in diesem Fall 5 bis 6 Monate lagen und im Anschluss an das Feedback kein weiterer Test zu bearbeiten war, für den die Anregungen hilfreich wären. Möglicherweise macht auch das Fehlen jeglicher Vergleichsinformation (die ja auch in der kriterialen Rückmeldung implizit enthalten ist, wenn alle möglichen Kompetenzstufen beschrieben werden) diese Rückmeldeform unattraktiv.

4. Zusammenfassung und Diskussion

In der ersten Teilstudie des Projekts Co^{2}CA konnten bereits substantielle Erkenntnisse zu den Kernfragestellungen gewonnen werden.

- In den hier untersuchten Realschulklassen dominiert nach Angaben der Lehrpersonen eine verbale Rückmeldekultur, verbunden mit verschiedenen lehrer- und notenzentrierten Beurteilungsformen. Partizipative Formen (Selbst- oder Peer-Evaluation) sind insgesamt selten, finden sich aber vergleichsweise häufig bei Lehrern und insbesondere Lehrerinnen, die angeben, sich gut mit diagnostischen Fragen auszuken-

nen. Diese Ergebnisse verweisen darauf, dass Beurteilungs- und Rückmeldekulturen von Lehrkraft zu Lehrkraft bzw. von Klasse zu Klasse variieren, sodass es sich lohnt, Ursachen und Folgen weiter zu prüfen.

- In Mehrebenenmodellen lassen sich tatsächlich Zusammenhänge mit Motivation und Leistung der Schüler identifizieren: lehrer- und notenzentrierte Beurteilungspraktiken gehen mit niedrigerer, eine aus Schülersicht wahrgenommene individuelle Bezugsnormorientierung der Lehrkraft hingegen mit höherer Motivation einher, während die diagnostische Kompetenz der Lehrperson (hier verstanden als hohe Treffsicherheit bei der Vorhersage von Schülerantworten) mit besseren Testleistungen der Schüler verbunden ist.
- Unterschiedliche Formen der Rückmeldung, die experimentell variiert wurden, wirkten sich erwartungsgemäß auf die Motivation der Schüler und auf deren Attribution der Testergebnisse aus, wobei insbesondere die kriteriale Rückmeldung auf der Basis von Kompetenzstufenmodellen signifikant bessere Effekte hatte als eine sozialnormorientierte Rückmeldung, wie sie in Schulklassen üblich ist.

Mit diesen Analysen, die hier nur in einem ersten, kursorischen Durchgang präsentiert werden konnten, beginnt das Projekt Co²CA seine Hauptfragestellung zu beantworten: die Frage nach dem Zusammenhang zwischen Leistungsbeurteilung/Feedback einerseits sowie Unterrichtsqualität und Lernergebnissen andererseits. Die Ergebnisse sind – auch wenn es sich hier nur um Querschnittsresultate handelt – kompatibel mit der Ausgangsthese des Projekts, dass eine argumentative, aktivierende, auf individuellen Bezugsnormen aufbauende Leistungsbeurteilung mit differenzierten (kriterialen bzw. prozessbezogenen) Rückmeldungen ein wichtiges Qualitätsmerkmal des Unterrichts darstellt. Für die Praxis ergibt sich als Schlussfolgerung, das Potential formativer Assessment-Praktiken stärker zu nutzen und bei summativen Testverfahren sorgfältig auf die Gestaltung der Rückmeldungen zu achten.

Literatur

Baker, E .(2007): The End(s) of testing. In: Educational Researcher 36, H. 6, S. 309–317.

Black, P.J./William, D. (1998): Assessment and Classroom Learning. In: Assessment in Education: Principles, Policy and Practice 5, H. 1, S. 7–74.

Brookhart, S.M. (1993): Teachers' grading practices: Meaning and values. In: Journal of Educational Measurement 30, S. 123–142.

Deci, E.L./Koestner, R./Ryan, R.M. (1999): A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. In: Psychological Bulletin 125, S. 627–668.

Deci, E.L./Schwartz, A.J./Sheinman, L./Ryan, R.M. (1981): An instrument to asses adults' orientations toward control versus autonomy with children: Reflections on intrinsic motivation and perceived competence. In: Journal of Educational Psychology 73, H. 5, S. 642–650.

Hattie, J./Timperley, H. (2007): The power of feedback. In: Review of Educational Research 77, H. 1, S. 81–112.

Heritage, M. (2007): Formative Assessment: What Do Teachers Need to Know and Do? In: Phi Delta Kappan 89, H. 2, S. 140–145.

Klieme, E. (2003): Benotungsmaßstäbe an Schulen: Pädagogische Praxis und institutionelle Bedingungen. Eine empirische Analyse auf der Basis der PISA-Studie. In: Döbert, H./von Kopp, B./Martini, R./Weiß, M. (Hrsg.): Bildung vor neuen Herausforderungen: historische Bezüge, rechtliche Aspekte, Steuerungsfragen, internationale Perspektiven. Neuwied/Kriftel: Luchterhand, S. 195–210.

Klieme, E./Pauli, C./Reusser, K. (2009): The Pythagoras Study. In: Janik, T./Seidel, T. (Hrsg.): The Power of Video Studies in Investigating Teaching and Learning in the Classroom. Münster: Waxmann, S. 137–160.

Koretz, D. (2008): Test-based Educational Accountability. Research Evidence and Implications. In: Zeitschrift für Pädagogik 54, H. 6, S. 777–790.

Pellegrino, J.W./Chudowsky, N./Glaser, R. (2001): Knowing what students know. The science and design of educational assessment. Washington, DC: National Academic Press.

Rakoczy, K./Klieme, E./Bürgermeister, A./Harks, B. (2008): The Interplay between Student Evaluation and Instruction: Grading and Feedback in Mathematics Classrooms. In: Zeitschrift für Psychologie/Journal of Psychology 216, H. 2, S. 111–124.

Sadler, D.R. (1989): Formative assessment and the design of instructional systems. In: Instructional Science 18, S. 119–144.

Schrader, F.-W. (^{3}2006): Diagnostische Kompetenz von Eltern und Lehrern. In: Rost, D.H. (Hrsg.): Handwörterbuch Pädagogische Psychologie. Weinheim/Basel: Beltz, S. 95–100.

Shavelson, R.J./Young, D.B./Ayala, C.C./Brandon, P.R./Furtak, E.M./Ruiz-Primo, M.A./Tomita, M.K./Yin, Y. (2008): On the Impact of Curriculum-Embedded Formative Assessment on Learning: A Collaboration between Curriculum and Assessment Developers. In: Applied Measurement in Education 21, S. 295–314.

Wilson, M. (2008): Cognitive Diagnosis Using Item Response Models. In: Zeitschrift für Psychologie/Journal of Psychology 216, S. 73–87.

Anschrift der Autor/innen

Prof. Dr. Eckhard Klieme, Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Schloßstraße 29, D-60486 Frankfurt a.M.
E-Mail: klieme@dipf.de

Anika Bürgermeister, M.A., Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Schloßstraße 29, D-60486 Frankfurt a.M.
E-Mail: buergermeister@dipf.de

Birgit Harks, Dipl. Psych., Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Schloßstraße 29, D-60486 Frankfurt a.M.
E-Mail: harks@dipf.de

Prof. Dr. Werner Blum, Universität Kassel; Fachbereich Mathematik, Heinrich-Plett-Str. 40, D-34132 Kassel
E-Mail: blum@mathematik.uni-kassel.de

Dr. Dominik Leiß, Leuphana Universität Lüneburg, Institut für Mathematik und ihre Didaktik, Scharnhorstr. 1 ,D-21335 Lüneburg
E-Mail: leiss@me.com

Dr. Katrin Rakoczy, Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Schloßstraße 29, D-60486 Frankfurt a.M.
E-Mail: rakoczy@dipf.de