Bolte, Claus; Machts, Nils; Möller, Jens; Wittchen, Sascha

Analyse diagnostischer Kompetenzen angehender Grundschullehrer*innen mit Studienfach Sachunterricht / Naturwissenschaften

Egger, Christina [Hrsg.]; Neureiter, Herbert [Hrsg.]; Peschel, Markus [Hrsg.]; Goll, Thomas [Hrsg.]: In Alternativen denken. Kritik, Reflexion und Transformation im Sachunterricht. Bad Heilbrunn : Verlag Julius Klinkhardt 2024, S. 15-26. - (Probleme und Perspektiven des Sachunterrichts; 34)

Claus Bolte, Nils Machts, Jens Möller und Sascha Wittchen

Analyse diagnostischer Kompetenzen angehender Grundschullehrer*innen mit Studienfach Sachunterricht / Naturwissenschaften

The diagnosis of student performance is one of the most important tasks of teachers. On the one hand, performance diagnoses form the basis for adaptive instructional planning, and on the other hand, performance-related assessments and prognoses (e.g., the recommendation for subsequent schooling) influence the students' educational pathways and thus their future. The analysis of teachers' diagnostic competencies, however, is very difficult due to the large number of variables relevant to practice. One possibility to obtain reliable information about diagnostic competencies of (prospective) teachers is the use of the Simulated Classroom (SCR), a digital tool for simulating complex and dynamic interactions during classroom discussions. In this article, we pursue the question: what is the state of diagnostic competencies of elementary student teachers majoring in integrated science? For this purpose, we developed the SCR Integrated Science (SCR Science) and asked N = 72 elementary student teachers of integrated science to conduct a classroom discussion with 12 simulated students in the SCR Science. Beforehand, the student teachers were given the task of judging the difficulties of the tasks used in the SCR Science. During the simulated classroom discussion, the task was to judge the qualities of individual student answers. At the end of the simulation, the students' overall performances were to be judged by the student teachers. In this contribution, we present the SCR Science and selected results.

1 Einleitung

Das akkurate Diagnostizieren der Leistungen von Schüler*innen gehört wohl zu den wichtigsten Aufgaben von Lehrer*innen (Kaiser & Möller 2017); denn einerseits sollten diese Diagnosen die Grundlage bei der Auswahl passender Unterrichtsmaterialien und / oder bei der Wahl geeigneter Maßnahmen zur Binnen-

doi.org/10.35468/6077-02

differenzierung bilden (Lorenz & Artelt 2009, 212); andererseits beeinflussen leistungsdiagnostische Beurteilungen und Prognosen (z. B. im Zuge der weiterführenden Schulempfehlung zum Ende der Grundschulzeit) maßgeblich die schulischen Bildungsgänge von Schüler*innen (Schrader 2013). Die Frage: *Wie ist es um die diagnostischen Kompetenzen von (angehenden) Grundschullehrer*innen mit Fach Sachunterricht / Naturwissenschaften bestellt?* ist daher von großer professionsbezogener Relevanz.

Die Analyse diagnostischer Kompetenzen (angehender) Lehrer*innen gestaltet sich hingegen wegen der Vielzahl der in der Praxis kaum zu kontrollierenden Variablen als sehr schwierig. Allein die Leistungsheterogenität einer Klasse übt einen starken Einfluss auf die Urteilsgenauigkeit aus und kann die Güte der Diagnose beeinträchtigen (Schrader 2013). Eine Möglichkeit, zuverlässige Aussagen über diagnostische Kompetenzen z. B. von (angehenden) Grundschullehrer*innen treffen zu können, stellt der Einsatz des Simulierten Klassenraums (kurz: SKR) dar – ein digitales Tool zur Simulation komplexer und dynamischer Interaktionen zwischen Lehrer*innen und Schüler*innen (Südkamp, Möller & Pohlmann 2008). Zur Beantwortung der oben allgemein formulierten Forschungsfrage haben wir den SKR-Nawi herangezogen, um eine weiter ausdifferenzierte und ökologisch valide(re) Version dieses Instruments zu entwickeln. Erprobt haben wir den SKR-Nawi in der Ausbildung angehender Grundschullehrer*innen. Im folgenden Abschnitt umreißen wir theoretische Grundlagen, auf denen die Entwicklung des SKR-Nawi basiert. Ferner stellen wir Kriterien vor, die das Konstrukt diagnostischer Kompetenzen näher beschreiben und unsere Forschungsfragen konkretisieren.

2 Theorie

2.1 Diagnostische Kompetenz und Urteilsgenauigkeit

Laut Schrader (2013, 154) bezieht sich der Terminus *Diagnostische Kompetenz* „auf die Fähigkeit, die im Lehrberuf anfallenden diagnostischen Aufgabenstellungen erfolgreich zu bewältigen, und auf die Qualität der dabei erbrachten Diagnoseleistungen". Anders, als diese recht allgemein gehaltene Definition suggerieren mag, handelt es sich bei der diagnostischen Kompetenz von Lehrer*innen keineswegs um ein unitäres Konstrukt, sondern um eine Vielzahl individueller Kompetenzen (Spinath 2005), die, wie Lorenz und Artelt (2009) zeigen konnten, vor allem fachspezifischer Natur sind.

Zu dieser Fülle eigenständiger fachspezifischer Kompetenzen von Lehrer*innen zählt in erster Linie die Fähigkeit, die *Leistungen von Schüler*innen* zutreffend zu beurteilen (Schrader 1989; Südkamp, Kaiser & Möller 2012). Allein diese Kompetenz bezieht verschiedene Beurteilungskriterien ein: Zum einen gilt es, jede einzelne Äußerung eines Schülers / einer Schülerin hinsichtlich ihrer fachlich-sachlichen und altersgemäßen Angemessenheit (Qualität) korrekt einzuordnen;

doi.org/10.35468/6077-02

wir bezeichnen diese Kompetenz „Kompetenz zur Beurteilung der Antwortqualität". Da Schüler*innen in der Regel keine druckreifen Antworten formulieren, die als eineindeutig und in Gesamtheit als fachlich richtig zu beurteilen wären, sind mit Blick auf dieses Beurteilungskriterium bereits qualitätsdifferenzierende Teilkompetenzen erforderlich (siehe weiter unten). Darüber hinaus sind die verschiedenen Äußerungen der Schüler*innen insgesamt und in Summe akkurat zu beurteilen („Beurteilung der Gesamt-Performanz der jeweiligen Schülerin / des jeweiligen Schülers"). Im Zuge der Festlegung der Schüler*innen-Leistungen in Tests und Klausuren dürfte dies keine allzu große Herausforderung darstellen; Leistungsbeurteilungen im Zuge komplexer und hoch dynamischer Unterrichtsgespräche stellen diesbezüglich die Lehrer*innen vor deutlich größere Probleme. Um die Qualität von Schüler*innen-Äußerungen differenziert zu bestimmen, ist zusätzlich der Schwierigkeitsgrad der Aufgabe bzw. die Komplexität und Tragweite der jeweiligen Schüler*innen-Antwort oder auch die Passung der jeweils vorgeschlagenen Problemlösung in die Beurteilung einzubeziehen. Dies setzt folgerichtig voraus, dass (angehende) Lehrer*innen auch über Fähigkeiten zur korrekten Einschätzung von *Aufgabenschwierigkeiten* verfügen (Schrader 1989; Südkamp, Kaiser & Möller 2012).

In der empirischen Bildungsforschung hat sich das Vorgehen bewährt, diagnostische Kompetenzen anhand von *Urteilsgenauigkeiten*, d. h. über das Maß der Übereinstimmungen von Beurteilungen im Vergleich zu den tatsächlich gezeigten Ausprägungen der zu beurteilenden Merkmale, zu bestimmen (Schrader 1989; Spinath 2005; Südkamp, Möller & Pohlmann 2008, Südkamp, 2020; Südkamp, Kaiser & Möller 2012). Schrader (1989) unterscheidet, unter Bezugnahme auf Cronbach (1955), drei *Komponenten der Urteilsgenauigkeit*: die *Rang-*, *Niveau-* und *Differenzierungskomponente*.

Die *Rangkomponente* quantifiziert, inwiefern die Beurteilungen der individuellen Leistungen von Schüler*innen die tatsächliche leistungsbezogene Rangfolge der Schüler*innen abbilden (Schrader 1989, 88).

Da Schüler*innen, die in einer bestimmten Klasse im interindividuellen Vergleich als vergleichsweise besonders leistungsstark oder -schwach zu beurteilen sind und diese Zuordnung jedoch in einer anderen Klasse deutlich anders ausfallen könnte, ist es ebenfalls wichtig zu wissen, inwiefern die individuellen Leistungsbeurteilungen von Lehrer*innen das mittlere Leistungsniveau einer Klasse abbilden. Die Genauigkeit der Abbildung des mittleren Leistungsniveaus einer Klasse wird anhand der *Niveaukomponente* beschrieben (Schrader 1989, 87).

Darüber hinaus ist es, mit Blick auf die Auswahl geeigneter binnendifferenzierender Maßnahmen, von Bedeutung, inwiefern die Beurteilung der individuellen Schüler*innen-Leistungen die Leistungsheterogenität der Klasse widerspiegelt. Die Genauigkeit der Abbildung der Leistungsheterogenität wird mit Hilfe der *Differenzierungskomponente* diskutiert (Schrader 1989, 87-88).

doi.org/10.35468/6077-02

Die *Urteilsgenauigkeit* im Allgemeinen und damit verbunden auch die drei *Komponenten der Urteilsgenauigkeit* im Speziellen werden in der Praxis durch eine Vielzahl von Variablen beeinflusst. Südkamp, Kaiser & Möller (2012) veranschaulichen mit Hilfe des *Heuristischen Modells der Akkuratheit diagnostischer Urteile von Lehrkräften* (s. Abb. 1), welche Variablen die Urteilsgenauigkeit unmittelbar und / oder mittelbar beeinflussen und inwiefern der Einfluss dieser Variablen von den Ausprägungen jeweils anderer Variablen abhängig sein könnte.

Das *Heuristische Modell der Akkuratheit diagnostischer Urteile von Lehrkräften* (Abb. 1) beschreibt zwei Gruppen von Merkmalen, die das Urteil einer Lehrperson beeinflussen; unterschieden werden: Lehrermerkmale (z. B. die diagnostischen Kompetenzen) und Urteilsmerkmale (z. B. die Spezifität des Urteils). Darüber hinaus beschreibt das Modell zwei weitere Gruppen von Merkmalen, die die Schülerleistungen beeinflussen. Hierbei handelt es sich um Schülermerkmale (z. B. die Ausprägung bestimmter Teilkompetenzen) und Testmerkmale (z. B. die Schwierigkeit unterschiedlicher Testaufgaben). Allen vier Gruppen von Merkmalen ist gemein, dass sie die Urteilsgenauigkeit von Lehrer*innen unmittelbar oder zumindest mittelbar beeinflussen.

Das Modell beschreibt außerdem mögliche Interdependenzen zwischen den Variablen, die Einfluss auf die Urteilsgenauigkeit ausüben (können). So ist es z. B. denkbar, dass Wechselwirkungen zwischen Lehrer- und Schülermerkmalen (beispielsweise bedingt durch Geschlechtsidentität oder ethnische Zugehörigkeit) auftreten. Vergleichbare Wechselwirkungen könnten zwischen den Urteils- und den Testmerkmalen bestehen. Unterscheiden sich beispielsweise die Spezifität der Merkmalsbeurteilung und die der Erfassung des Merkmals durch einen Test, so wirkt sich dies negativ auf die Urteilsgenauigkeit aus (Südkamp et al. 2012).

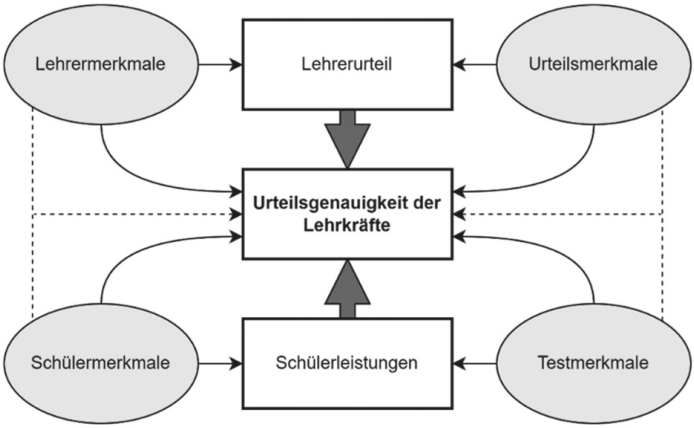

Abb. 1: Heuristisches Modell der Akkuratheit diagnostischer Urteile von Lehrkräften von Südkamp, Kaiser & Möller (2012, 756).

doi.org/10.35468/6077-02

2.2 Der Simulierte Klassenraum

Aufgrund der Vielzahl der in empirischer Forschung kaum zu kontrollierenden Variablen, die die Urteilsgenauigkeit in der Praxis beeinflussen, und angesichts der zahlreichen potentiellen Wechselwirkungen zwischen diesen, liegt es nahe, auf Simulationen diagnostischer Situationen zurückzugreifen, da diese ein hohes Ausmaß an Variablenkontrolle ermöglichen (Machts, Chernikova, Jansen, Weidenbusch, Fischer & Möller 2023).

Ein Bespiel einer solchen Simulation diagnostischer Aktivitäten ist der *Simulierte Klassenraum* (Fiedler, Walther, Freytag & Plessner 2002; Südkamp, Möller & Pohlmann 2008; Bolte, Köppen, Möller & Südkamp 2011; Kaiser et al. 2017; Wittchen, Bolte, Machts & Möller 2022). Im *Simulierten Klassenraum* (*SKR*) übernehmen die Teilnehmer*innen die Rolle von Lehrer*innen, die simulierten Unterrichtsgespräche mit den virtuell animierten Schüler*innen führen, indem sie diesen Fragen stellen oder Aufgaben geben. Die simulierten Schüler*innen beantworten die an sie gerichteten Fragen und Aufgaben entsprechend ihrer vorab festgelegten Leistungsparameter. Nach Beendigung des Unterrichtsgespräches werden die Teilnehmer*innen gebeten, die Leistungen der simulierten Schüler*innen zu beurteilen. Basierend auf den Leistungsbeurteilungen und den "empirisch erbrachten" Leistungen der simulierten Schüler*innen können nun die *Urteilsgenauigkeiten* der Teilnehmer*innen berechnet werden. Die auf diesem Wege bestimmten Ergebnisse erlauben wiederum Rückschlüsse auf die diagnostischen Kompetenzen der Teilnehmer*innen.

In der Vergangenheit konnte bereits gezeigt werden, dass sich der *SKR* sowohl durch eine hohe interne Validität als auch durch eine vergleichsweise hohe ökologische Validität auszeichnet (Kaiser et al. 2017).

2.3 Forschungsfragen

Vor dem Hintergrund der hier dargelegten Überlegungen verfolgen wir in diesem Beitrag die folgenden theoriegeleiteten Forschungsfragen: *Wie akkurat beurteilen Studierende des Grundschullehramts mit Studienfach Sachunterricht/Naturwissenschaften…*

- die Schwierigkeit der jeweiligen Aufgaben für SKR-Naturwissenschaften?
- die fachliche Qualität der Aussagen der (simulierten) Schüler*innen im SKR-Nawi?
- die von den simulierten Schüler*innen im Verlauf des Unterrichtsgesprächs gezeigten Leistungen?

doi.org/10.35468/6077-02

3 Methode

Zur Beantwortung der Forschungsfragen haben wir den *SKR* adaptiert und den *Simulierten Klassenraum Integrierte Naturwissenschaften 5/6* (*SKR-Nawi*), eine fach- und schulformspezifische Version des *SKR*, entwickelt. Im *SKR-Nawi* übernehmen die Versuchsteilnehmer*innen die Rolle von Nawi-Lehrer*innen, die *Aufgabenschwierigkeiten*, einzelne *Aussagen von Schüler*innen* sowie die *Gesamtleistung simulierten Schüler*innen* nach Ablauf eines (in dieser Studie 20-minütigen) Unterrichtsgesprächs beurteilen sollen.

3.1 Aufgaben und Schülerantworten für den SKR-Nawi

Für den *SKR-Nawi* haben wir 45 Aufgaben entwickelt, die sich für den Integrierten Naturwissenschaftlichen Anfangsunterricht der Jahrgangsstufe 5/6 eignen. Die Konstruktion der Aufgaben erfolgt entsprechend der Vorgaben der KMK (2013) sowie des Rahmenlehrplans Naturwissenschaften Jahrgangsstufe 5/6 (SenBJF & MinBJS 2015). Darüber hinaus haben wir uns bei der Entwicklung der Aufgaben an den fachlichen Inhalten und thematischen Schwerpunkten des Schulbuches Natur und Technik – Naturwissenschaften 5/6 von Bresler et al. (2017) orientiert. Um eine möglichst gleichmäßige Repräsentation der drei "klassischen naturwissenschaftlichen Disziplinen" zu gewährleisten, entfallen jeweils 15 Aufgaben auf die Themenbereiche *Aggregatzustände & Teilchenmodell* (Chemie/Physik), *Sinne & Messen* (Biologie/Physik) und *Körper & Gesundheit* (Biologie/Chemie). Von diesen 15 Aufgaben pro Themenfeld bilden jeweils fünf Aufgaben die Schwierigkeitsniveaus *leicht*, *mittelmäßig anspruchsvoll* und *schwierig* ab. Die Abbildung der Schwierigkeitsniveaus erfolgte konzeptionell in Anlehnung an die Leitlinien der KMK (2013) und auf Basis der Arbeiten von Kauertz (2008) zur Abhängigkeit von Aufgabenschwierigkeiten sowohl von der Anzahl der zu verknüpfenden Elemente als auch von der Qualität der Verknüpfung vorgenommen. Für jede der insgesamt 45 Aufgaben wurden drei mögliche Schüler*innen-Antworten konzipiert, die die Antwortqualitäten *richtig*, *in Teilen richtig/unvollständig* und *falsch* abbilden sollen.

3.2 Der Simulierte Klassenraum Naturwissenschaften 5/6

Vor Beginn des simulierten Unterrichtsgesprächs werden die Teilnehmer*innen gebeten, die Aufgaben, die sie später für die Interaktionen mit den simulierten Schüler*innen verwenden werden, hinsichtlich ihrer *Aufgabenschwierigkeiten* zu beurteilen. Anschließend führen die Teilnehmer*innen der hier berichteten Studie je zwei zwanzigminütige Unterrichtsgespräche mit denselben 12 simulierten Schüler*innen durch (Abb. 2).

doi.org/10.35468/6077-02

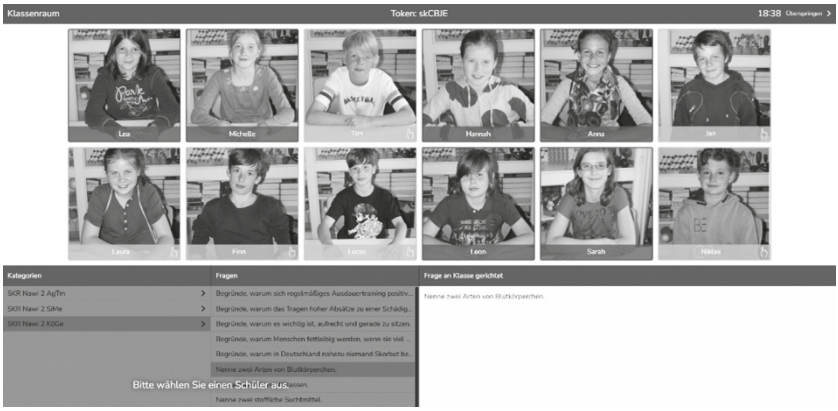

Abb. 2: Benutzeroberfläche des Simulierten Klassenraum Naturwissenschaften (SKR-Nawi)

Die simulierten Interaktionen sind so gestaltet, dass die Teilnehmer*innen Fragen auswählen, die die aufgerufenen Schüler*innen entsprechend ihrer voreingestellten Leistungsparameter *richtig, in Teilen richtig bzw. unvollständig oder falsch* beantworten. Die Teilnehmer*innen werden nach jeder Interaktion gebeten, die *Antwortqualität* (s. o.) zu beurteilen. Am Ende jedes Unterrichtsgespräches werden die Teilnehmer*innen aufgefordert, die *Gesamtleistungen* der simulierten Schüler*innen in Form prozentualer Lösungsanteile zu beurteilen.

4 Ergebnisse und Diskussion

72 Studierende des Grundschullehramts mit Studienfach Sachunterricht / Naturwissenschaften haben an der Studie teilgenommen (n = 50 Bachelorstudierende und n = 22 Masterstudierende).

Die Ergebnisse zeigen, dass die Beurteilungen der *Aufgabenschwierigkeiten* sowohl den Bachelor- ($\bar{K}_W = 0.47$) als auch den Masterstudierenden ($\bar{K}_W = 0.45$) mit *moderater* Genauigkeit gelungen sind (vgl. Landis & Koch, 1977).

Auch die Beurteilungen der *Antwortqualitäten* während der Interaktionen mit den simulierten Schüler*innen sind von Bachelor- ($\bar{K}_{W.1} = 0.44$; $\bar{K}_{W.2} = 0.55$) und Masterstudierenden ($\bar{K}_{W.1} = 0.45$; $\bar{K}_{W.2} = 0.51$) mit *moderater* Genauigkeit (vgl. Landis & Koch, 1977) vorgenommen worden.

Die Genauigkeit der Beurteilungen der Gesamtleistungen der simulierten Schüler*innen differiert hinsichtlich der drei *Beurteilungskomponenten* nach Schrader (1989); die Ergebnisse sind Tabelle 1 zu entnehmen.

Die *Rangkomponente* wird über den personenbezogenen Zusammenhang zwischen den Beurteilungen der Schüler*innen und deren tatsächlich gezeigten

doi.org/10.35468/6077-02

Leistungen in Form der relativen Häufigkeit vollständig korrekter Antworten bestimmt. Im Zuge dieser Analysen wird die Pearson Korrelation verwendet, bei der theoretisch ein maximaler Genauigkeitswert von $r = 1$ erzielt werden kann. Die Werte der *Rangkomponente* zeigen, dass es den Studierenden nur in geringfügigem Maß gelungen ist, die Leistungsvarianz der simulierten Schüler*innen akkurat einzuschätzen. Die Bachelorstudierenden urteilten nach dem zweiten Durchlauf signifikant genauer ($p < 0.05$) als nach dem ersten Durchlauf. Diese Steigerung ist den Masterstudierenden nicht zu bescheinigen.
Die *Niveaukomponente* wird über die personenbezogene Differenz der mittleren Beurteilungen und der mittleren Leistungen aller Schüler*innen einer Klasse berechnet. Im Zuge der Berechnung wird die Mittelwertdifferenz bestimmt; der theoretisch bestmöglich zu erzielende Wert ist gleich null ($\Delta M = 0$). Werte kleiner null bringen strengere, Werte größer null mildere Leistungseinschätzungen zum Ausdruck. Die Ergebnisse zur Bestimmung der *Niveaukomponente* zeigen, dass beide Teilstichproben das mittlere Leistungsniveau der Klasse deutlich (zwischen 13 und 24 Prozentpunkten) überschätzen. Die inferenzstatistischen Vergleiche der Beurteilungen beider Gruppen belegen, dass die Bachelorstudierenden das mittlere Leistungsniveau der simulierten Klasse für beide Durchläufe akkurater bewertet haben ($p < 0.05$) als die Masterstudierenden (s. Tab. 1).

Tab. 1: Mittelwerte und *Standardabweichungen* der drei Komponenten der Leistungsbeurteilung – in Anlehnung an Schrader (1989).

Stichprobe	n	Rangkomponente: M_{RK} (*SD*)		Niveaukomponente: M_{NK} (*SD*)		Differenzierungskomponente M_{DK} (*SD*)	
Durchgang		DG 1	DG 2	DG 1	DG 2	DG 1	DG 2
Bachelor	50	0.14 (*0.36*)	0.27 (*0.29*)	0.14 (*0.14*)	0.13 (*0.20*)	0.47 (*0.22*)	0.44 (*0.21*)
Master	22	0.17 (*0.31*)	0.17 (*0.25*)	0.22 (*0.14*)	0.24 (*0.11*)	0.47 (*0.26*)	0.43 (*0.30*)
Gesamt	72	0.15 (*0.34*)	0.24 (*0.28*)	0.16 (*0.14*)	0.16 (*0.18*)	0.47 (*0.23*)	0.44 (*0.24*)

Die *Differenzierungskomponente* wird über den personenbezogenen Abgleich von Beurteilungsstreuung und Leistungsstreuung über alle Schüler*innen einer Klasse bestimmt. Dabei wird der Quotient aus den jeweiligen Standardabweichungen gebildet, bei dem der theoretisch bestmöglich zu erzielende Wert der Zahl eins entspricht ($SD_{\text{Beurteilung}} : SD_{\text{Leistung}} = 1$). Werte kleiner eins sprechen für eine Tendenz zur Mitte und (besonders) hohe Werte für extreme(re) Einschätzungen. Den Werten der *Differenzierungskomponente* zur Folge haben die Proband*innen

doi.org/10.35468/6077-02

die von den simulierten Schüler*innen gezeigte Leistungsheterogenität zwischen 43 % und 47 % der Fälle korrekt eingeschätzt. Die t-Test-Analysen weisen keine statistisch signifikanten Unterschiede zwischen den Teilstichproben und den zwei Testdurchläufen aus (s. Tab. 1).

5 Interpretation der Ergebnisse – Fazit

Insgesamt belegen die Ergebnisse dieser Studie, dass die hier untersuchten Studierenden des Grundschullehramts mit Fach Sachunterricht / Naturwissenschaften hinsichtlich ihrer diagnostischen Kompetenzen dringend weiterer Förderung bedürfen. Dieses Fazit zeigt sich bereits in den Ergebnissen zur Beurteilung der *Aufgabenschwierigkeiten.* Offensichtlich wurde die Fähigkeit zur systematisch angelegten Analyse von Aufgabenschwierigkeiten mit den Studierenden nicht oder nicht umfassend genug erarbeitet und eingeübt. Zwei Parameter spielten in der Konstruktion der Aufgaben eine schwierigkeitsbestimmende Rolle; zum einen die Anzahl korrekter Aufzählungen und sachgemäßer Verknüpfung von Antwortelementen und zum anderen das Anforderungsniveau des verwendeten Operators (benenne, beschreibe, erkläre). Fehleinschätzungen können daher rühren, dass subjektive und fachinhaltlich motivierte Schwierigkeitseinschätzungen vorgenommen wurden; im Sinne: physikalisch anmutende Aufgaben könnten als schwerer erachtet worden sein als Aufgaben aus dem Bereich der Biologie.
Größere Sorgen bereiten uns die lediglich als moderat zu beurteilenden Fähigkeiten der Proband*innen, die *Antwortqualitäten der simulierten Schüler*innen* korrekt einschätzen zu können. Dass im Mittel strenggenommen nur jede zweite Schüler*innen-Antwort korrekt beurteilt und den simulierten Schüler*innen entsprechend zurückgemeldet wurde, gibt zu denken. Da die Antworten der simulierten Schüler*innen aus den Wissensbereichen der Jahrgangsstufen 5 / 6 stammen und die Antworten der simulierten Schüler*innen bildungssprachlich korrekt und wohl verständlich formuliert gewesen sind, scheint sich hier ein ernstzunehmendes Qualifikationsproblem zu offenbaren. Sucht man die Ursachen für dieses Problem nicht ausschließlich in den "ausbaufähigen" fachwissenschaftlichen Kompetenzen seitens der Studierenden, so stellt sich schon die Frage: Inwiefern werden in grundlegenden fachwissenschaftlichen Lehrveranstaltungen des Studiengangs die Basiskonzepte thematisiert und erfolgreich vermittelt, deren professionsbezogene Beherrschung für angehende Grundschullehrer*innen in der Praxis des Sach- bzw. des naturwissenschaftlichen Unterrichts als zwingend notwendig erachtet wird? Mit anderen Worten: Wird – wenn auch in guter Absicht – zu viel vom Falschen gelehrt und von den Studierenden schlussendlich nicht genügend sinnstiftend gelernt?
Mit Blick auf die Ergebnisse zu den drei zentralen Komponenten akkurater Leistungsbeurteilungen, *der Rang-, Niveau- und Differenzierungskomponente*, dürfen

doi.org/10.35468/6077-02

und wollen wir mit den Teilnehmer*innen dieser Untersuchung nicht zu streng ins Gericht gehen. Die Aufgabenstellung in dieser experimentell angelegten Studie war und ist herausfordernd. Die simulierten Schüler*innen waren eben keine lebhaften Kinder mit eindrucksvollen Verhaltensweisen und einprägsamen Charaktereigenschaften. Außerdem ist zu bedenken, dass es sich für die Proband*innen um eine völlig ungekannte Schulklasse gehandelt hat, und dass die Leistungsbeurteilung schon nach einer verhältnismäßig kurzen Zeit zu erfolgen hatte. Auch die Testsituation an sich könnte zu Verunsicherungen geführt und so die Befunde fehlerhaft beeinflusst haben.

Diese Bedenken und Limitationen sollen aber die Ergebnisse der Studie nicht per se in Frage stellen. Die in diesem Beitrag zur Diskussion gestellten Ergebnisse decken sich mit Befunden vorangegangener Studien (siehe Kap. 2). Sie regen, wie wir gerade versucht haben deutlich zu machen, zu (selbst-)kritischen Reflexionen an, u. a. wie das Grundschullehramtsstudium mit Schwerpunkt Sachunterricht / Naturwissenschaften sowohl fach- als auch berufswissenschaftlich optimiert werden könnte. Außerdem sind die Anlässe zur kritischen Selbstreflexion auf Seiten der Studienteilnehmer*innen u. E. nicht zu unterschätzen, die wir durch die gemeinsame Diskussion der Untersuchungsergebnisse mit den Seminarteilnehmer*innen eröffnen.

Im angeleiteten Selbstversuch zu erleben, wie schwer und herausfordernd es ist, das Leistungsniveau von Schulklassen *nicht* zu über- und die Leistungsheterogenität von Lerngruppen *nicht* zu unterschätzen, ist professionsbezogen betrachtet von hohem Wert. Ebenso wertvoll und lehrreich ist es u. E., wenn (angehende) Grundschullehrer*innen am eigenen Leib erfahren, wie häufig ihnen unzutreffende oder zumindest zum Teil unzutreffende Leistungsrückmeldungen an Schüler*innen unterlaufen. Blinde Flecke und vorhandene Unzulänglichkeiten zu identifizieren, aber vor allem Optimierungspotenziale aufzudecken, um diese entsprechend in Angriff zu nehmen, sind wichtige Meilensteine im Prozess kontinuierlich betriebener Professionalisierung. Die Anwendung des Tools der Simulierte Klassenraum Naturwissenschaften 5 / 6 in Forschung und Lehre stellen hierfür gewinnbringende Orientierungshilfen in Aussicht.

Literatur

Bresler, S, Heepmann, B., Kuck, C., Lichtenberger, J. & Rau, V. (2017): Natur und Technik – Naturwissenschaften. Berlin/Brandenburg 5./6. Schuljahr. Schülerbuch. Neubearbeitung. Berlin.

Bolte, C., Köppen, G., Möller, J. & Südkamp, A. (2011): Kompetenzdiagnostik im virtuellen naturwissenschaftlichen Unterricht. In: Höttecke, D. (Hrsg.): Naturwissenschaftliche Bildung als Beitrag zur Gestaltung partizipativer Demokratie. Münster, S. 146-148.

Cronbach, L. J. (1955): Processes affecting scores on "understanding of others" and "assumed similarity". In*:* Psychological Bulletin, 52, No. 3, 177-193. https://doi.org/10.1037/h0044919.

doi.org/10.35468/6077-02

Fiedler, K., Walther, E., Freytag, P. & Plessner, H. (2002): Judgment Biases in a Simulated Classroom - A Cognitive-Environmental Approach. In: Organizational Behavior and Human Decision Processes, 88, No. 1, 527-561. https://doi.org/10.1006/obhd.2001.2981.

Landis, J. R. & Koch, G. G. (1977): The Measurement of Observer Agreement for Categorical Data. In: Biometrics, 33, No. 1, 159-174. https://doi.org/10.2307/2529310.

Lorenz, C. & Artelt, C. (2009): Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. In: Zeitschrift für Pädagogische Psychologie, 23, Nr. 34, 211–222. https://doi.org/10.1024/1010-0652.23.34.211.

Kaiser, J. & Möller, J. (2017): Diagnostische Kompetenz von Lehramtsstudierenden. In: Gräsel, C. & Trempler, K. (Hrsg.): Entwicklung von Professionalität pädagogischen Personals. Wiesbaden, S. 55-75. https://doi.org/10.1007/978-3-658-07274-2_4.

Kaiser, J., Südkamp, A. & Möller, J. (2017): The Effect of Student Characteristics on Teachers' Judgement Accuracy: Disentangling Ethnicity, Minority Status, and Achievement. In: Journal of Educational Psychology, 109, 871-888. https://doi.org/10.1037/edu0000156.

KMK – Sekretariat der ständigen Kultusministerkonferenz der Länder in der Bundesrepublik Deutschland (2013): Operatoren für die naturwissenschaftlichen Fächer (Physik, Biologie, Chemie) an den Deutschen Schulen im Ausland. https://www.kmk.org/fileadmin/ Dateien/pdf/Bildung/Auslandsschulwesen/Kerncurriculum/Auslandsschulwesen-Operatoren-Naturwissenschaften-02-2013.pdf [05.2023].

Kauertz, A. (2008): Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben. Logos.

Machts, N., Chernikova, O., Jansen, T., Weidenbusch, M., Fischer, F. & Möller, J. (in press): Categorization of Simulated Diagnostic Situations and the Salience of Diagnostic Information: Conceptual Framework. In: Zeitschrift für Pädagogische Psychologie, 1-14. https://doi.org/10.1024/1010-0652/a000364.

Schrader, F.-W. (1989): Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts. Frankfurt am Main.

Schrader, F.-W. (2013): Diagnostische Kompetenz von Lehrpersonen. In: Beiträge zur Lehrerbildung, 31, Nr. 2, 154-165.

SenBJF & MBJS: Senatsverwaltung für Bildung, Jugend & Familie, & Ministerium für Bildung, Jugend & Sport. (2015): Rahmenlehrplan Teil C Naturwissenschaften Jahrgangsstufen 5/6. https://bildungsserver.berlin-brandenburg.de/fileadmin/bbb/unterricht/rahmenlehrplaene /Rahmenlehrplanprojekt/…2015_11_16_web.pdf [05.2023].

Spinath, B. (2005): Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. In: Zeitschrift für pädagogische Psychologie, 19, Nr. 1/2, 154-165. https://doi.org/10.1024/1010-0652.19.12.85.

Südkamp, A. (2010): Diagnostische Kompetenz: Zur Genauigkeit der Beurteilung von Schülerleistungen durch Lehrkräfte. Dissertation zur Erlangung des Doktorgrades der Philosophischen Fakultät der Christian-Albrechts-Universität zu Kiel.

Südkamp, A., Möller, J. & Pohlmann, B. (2008): Der Simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz. In: Zeitschrift für pädagogische Psychologie, 22, Nr. 34, 261-276. https://doi.org/10.1024/1010-0652.22.34.261.

Südkamp, A., Kaiser, J. & Möller, J. (2012): Accuracy of teachers' judgements of students' academic achievement: A meta-analysis. In: Journal of Educational Psychology, 104, 743-762.

Wittchen, S., Bolte, C., Machts, N. & Möller, J. (2022): Erfassung diagnostischer Kompetenzen Lehramtsstudierender des Faches Chemie. In: Habig, S. (Hrsg.): Unsicherheit als Element von naturwissenschaftsbezogenen Bildungsprozessen. Universität Duisburg-Essen, S. 444-447.

doi.org/10.35468/6077-02

Autorenangaben

Prof. Dr. Claus Bolte
Didaktik der Chemie
Freie Universität Berlin
claus.bolte@fu-berlin.de

Dr. Nils Machts
https://orcid.org/0000-0001-6829-3602
Institut für Pädagogisch-Psychologische Lehr- und Lernforschung
Christian-Albrechts-Universität zu Kiel
nmachts@ipl.uni-kiel.de

Sascha Wittchen
Freie Universität Berlin (bis 30.06.2023)

doi.org/10.35468/6077-02