Forschungsmonitor Schule
Rezensionen für die Praxis

Heinz Sander

Rezension zu

Schult, J. & Lindner, M. A. (2018). Diagnosegenauigkeit von Deutschlehrkräften in der Grundschule: Eine Frage des Antwortformats? *Zeitschrift für Pädagogische Psychologie, 32*(1–2), 75–87.

Kommentierter Kurzbefund

Lehrkräfte müssen die Schwierigkeit von Aufgabenmaterialien und die Klassen- bzw. Schülerleistungen, die bei der Bearbeitung zu erwarten sind, zutreffend einschätzen, um passende pädagogische Entscheidungen treffen zu können. Die Akkuratheit ihrer Diagnose hängt dabei von verschiedenen Faktoren ab, wobei dem Antwortformat des Aufgabenmaterials bisher wenig Beachtung geschenkt wurde.

Schult und Lindner untersuchen die Diagnoseakkuratheit von Lehrkräften zum einen bei Aufgaben mit geschlossenem Antwortformat, bei dem die Antwortmöglichkeiten vorgegeben sind, und zum anderen bei Aufgaben mit offenem Antwortformat, bei dem die Antworten selbst formuliert werden müssen. Hierfür stellen sie Einschätzungen von Lehrkräften zur Schwierigkeit von Aufgaben aus den VERA 3-Tests den tatsächlichen Testergebnissen ihrer Schülerinnen und Schüler gegenüber.

Im Ergebnis werden Lösungshäufigkeiten von geschlossenen Antwortformaten eher unterschätzt, diejenigen offener Antwortformate eher überschätzt. Bei offenen Antwortformaten können Lehrkräfte die Schwierigkeiten der verschiedenen Aufgaben zutreffender zueinander ins Verhältnis setzen. Als mögliche Ursachen diskutieren Schult und Lindner die Vernachlässigung der Ratewahrscheinlichkeit bei geschlossenen Aufgabenformaten und die andernorts nachgewiesene Unterschätzung von Schülerleistungen durch Lehrkräfte bei sehr leichten Aufgaben, da die geschlossenen VERA 3-Aufgaben im Schnitt besser gelöst wurden. Angesichts u. a. des diagnostischen Potenzials geschlossener Antwortformate v. a. für die Lernverlaufsdiagnostik plädieren sie für eine verstärkte Auseinandersetzung von Lehrkräften mit den Eigenschaften und Einsatzmöglichkeiten unterschiedlicher Aufgabenformate.

Diese Empfehlung liefert einen plausiblen Ansatzpunkt für die Verbesserung der Diagnosefähigkeit von Lehrkräften, allerdings ist sie nicht direkt aus den Untersuchungsergebnissen ableitbar. Zudem verlieren Testformate, die auf einem Richtig-falsch-Muster basieren, im Verlauf der Schullaufbahn zumindest in sprachlichen Fächern zugunsten unterschiedlich komplexer offener Formate an Bedeutung, weshalb zur weitergehenden Untersuchung des Einflusses des Aufgabenformats auf die Diagnoseakkuratheit vollkommen neue Instrumente entwickelt werden müssten.

Hintergrund

Zu den zentralen Anforderungen an Lehrkräfte zählt die Diagnosefähigkeit. Dazu gehört eine Einschätzung

- der Anforderungen, die sich aus den gestellten Aufgaben für Schülerinnen und Schüler ergeben, sowie
- der tatsächlich erbrachten Schülerleistung.

Das Ausmaß, in dem sich die Einschätzungen der Aufgaben und die erbrachten Leistungen entsprechen, bezeichnen Schult und Lindner als Diagnoseakkuratheit. Zwar gibt es hierzu bereits eine umfangreiche Forschung, doch differieren die dabei erzielten Ergebnisse so weit, dass die Autorin und der Autor die Suche nach Faktoren vorschlagen, welche situations- oder bereichsbedingt Einfluss auf die Diagnoseakkuratheit haben können.

In ihrer Arbeit gehen sie deshalb möglichen Einflüssen des Aufgabenformats auf die Diagnoseakkuratheit nach, zumal zu diesem Punkt ein Forschungsdesiderat besteht. Dazu unterscheiden sie hinsichtlich der schriftlichen Erfassung von Schulleistungen zwei grundsätzliche Formate:

1. Bei geschlossenen Formaten werden dem Schulkind für eine Frage mehrere mögliche Antworten präsentiert, aus denen es eine oder mehrere richtige auszuwählen hat. Der Multiple-Choice-Test dürfte die bekannteste Variante dieses Formats sein. Der effizienten Erfassung der Leistung stehen nicht nur inhaltliche Limitierungen gegenüber (Kreativität kann nicht erfasst werden). Auch Ergebnisverfälschungen sind möglich, da nicht ausgeschlossen werden kann, dass Ergebnisse geraten wurden.
2. Im Fall offener Formate hingegen sollen Schülerinnen und Schüler die Antwort eigenständig formulieren. Die Autorin und der Autor nehmen aufgrund „informeller Rückmeldungen“ an, dass offene Formate eine höhere Akzeptanz auf der Lehrkraftseite haben als die geschlossenen Formate und deshalb auch häufiger im Unterricht eingesetzt werden. Allerdings sei nicht auszuschließen, dass bei den offenen Formaten neben der Kompetenz, die eigentlich erfasst werden soll, auch Schreib- und Ausdrucksfähigkeit berücksichtigt würden und so Ergebnisse verzerrt würden.

Um die Diagnoseakkuratheit angemessen untersuchen zu können, unterscheiden Schult und Lindner drei Komponenten: 1. Niveaukomponente, 2. Differenzierungskomponente und 3. Rangordnungskomponente. Zur konkreten mathematischen Fassung der Komponenten vergleiche das Kapitel „Design“.

1. Bei der Niveaukomponente wird festgestellt, wie stark die Einschätzung von Seiten der Lehrenden zum Schwierigkeitsgrad einer Aufgabe durch die tatsächlich von den Schülerinnen und Schülern erbrachten Ergebnisse bestätigt wird. Da sich hierbei Unter- und Überschätzung von Aufgaben gegenseitig aufheben können, ermitteln die Autorin und der Autor auch den mathematischen Betrag der Abweichungen, den sie als „Urteilsfehler“ bezeichnen.
2. Die Differenzierungskomponente ergibt sich aus dem Verhältnis der Streuung der Lehrkrafturteile zu den Aufgaben gegenüber der Streuung der tatsächlichen Lösungshäufigkeiten.
3. Die Rangordnungskomponente wiederum entsteht durch den Vergleich der von der Lehrkraft erwarteten Lösungshäufigkeiten mit der tatsächlichen Rangfolge der Lösungshäufigkeiten in der Klasse.

Zwar gibt es zu diesen Komponenten bereits Untersuchungsergebnisse, die von der Autorin und dem Autor auch referiert werden. So wird etwa in den bisherigen Studien gezeigt, dass bei der Niveaukomponente die Aufgabenschwierigkeit regelhaft überschätzt wird. Eine Untersuchung des konkreten Einflusses des Aufgabenformats auf die Diagnoseakkuratheit steht allerdings noch aus.

Auf der Basis dieser Überlegungen fragen Schult und Lindner danach, ob die beiden oben skizzierten Aufgabenformate Konsequenzen für die Diagnoseakkuratheit haben. Eine besondere Rolle für die Formulierung von Erwartungen spiele dabei, dass geschlossene Formate im Unterricht seltener eingesetzt würden als offene, wodurch Lehrpersonen eine geringere Vertrautheit mit den geschlossenen Formaten hätten.

Die vier Hypothesen des Autors und der Autorin sind:

- Die Niveaukomponente soll bei den offenen Formaten einen größeren Wert annehmen als bei den geschlossenen.
- Der Urteilsfehler soll bei geschlossenen Formaten größer sein als bei offenen.
- Die Differenzierungskomponente soll eine schlechtere Differenzierung bei den geschlossenen Formaten anzeigen als bei offenen. In diesem Rahmen erwarten Schult und Lindner, dass die Streuung der Lehrkrafturteile größer als die Streuung der tatsächlichen Lösungshäufigkeiten ist.
- Die Rangordnungskomponente soll bei den geschlossenen Formaten geringer ausfallen als bei offenen.

Design

Zur Hypothesenprüfung ziehen Schult und Lindner die Ergebnisse der Lesekompetenzmessung im Fach Deutsch des regelmäßig im dritten Schuljahr bundesweit durchgeführten schriftlichen Leistungstests VERA 3 heran. Aus diesem Ergebniskorpus wählen die Autorin und der Autor die Ergebnisse der Jahre 2012 bis 2016 von 973 Klassen (bereinigt um 13 Klassen mit statistischen Auffälligkeiten) mit 17.586 teilnehmenden Schülerinnen und Schülern aus Baden-Württemberg aus.

Den Deutschlehrkräften der teilnehmenden Klassen wurden in der Woche vor der Durchführung des VERA 3-Tests die Aufgaben vorgelegt. Sie sollten dann angeben, wie vielen Schülerinnen und Schülern die Lösung der jeweiligen Aufgabe gelingen werde. Damit lag die Lehrerdiagnose vor, welche mit den realen Ergebnissen des kurz danach durchgeführten VERA-3-Tests verglichen werden konnte. Da diese Tests sieben bis 13 Aufgaben jeweils des offenen und geschlossenen Formats enthielten, ergab sich die Möglichkeit der Differenzierung nach Formaten.

Es wurden drei Komponenten der Diagnoseakkuratheit für die beiden Antwortformate berechnet:

- Die Niveaukomponente wurde berechnet, indem klassenweise von der erwarteten Zahl der richtigen Lösungen (dividiert durch die Zahl der Kinder einer Klasse) die tatsächliche Zahl der richtigen Lösungen (dividiert durch die Zahl der Kinder einer Klasse) abgezogen wurde. Diese Zahlen lagen auch der Ermittlung des Urteilsfehlers zugrunde.
- Bei der Feststellung der Differenzierungskomponente wurden die Standardabweichungen der Lehrkräfteeinschätzungen klassenweise über die unterschiedlichen Aufgaben hinweg durch die Standardabweichungen der tatsächlichen Lösungen dividiert.
- Die Rangordnungskomponente wurde klassenweise als Produkt-Moment-Korrelation der Lehrkrafteinschätzung und der empirischen Lösungshäufigkeit ermittelt.

Alle Berechnungen wurden jeweils separat für jedes Erhebungsjahr und für jedes der Antwortformate durchgeführt.

Ergebnisse

Niveaukomponente

Bei der Überprüfung der Niveaukomponente zeigt sich signifikant für alle fünf Jahrgänge, dass Lehrkräfte die Lösungshäufigkeiten bei offenen Formaten über- und bei geschlossenen Formaten unterschätzen. Der Befund entspricht der Hypothese von Schult und Lindner, steht aber insofern im Widerspruch zu älteren Forschungsarbeiten, als dass diese eine Überschätzung für beide Formate feststellten.

Allerdings ergibt sich für den Urteilsfehler nicht das von der Autorin und dem Autor erwartete Ergebnis: Nur für eines der fünf Jahre liegt ein signifikantes Ergebnis vor, wobei der Betrag der Verschätzung beim geschlossenen Format um ein Geringes kleiner ist als beim offenen.

Die Unterschätzung der Leistungen bei geschlossenen Aufgabenformaten könnte nach Ansicht des Autors und der Autorin möglicherweise u. a. auf eine fehlende Berücksichtigung der Ratewahrscheinlichkeit beim geschlossenen Format oder eine Unterschätzung der Leistung von Schülerinnen und Schülern bei den insgesamt etwas leichter zu lösenden geschlossenen Formaten zurückzuführen sein, da Lehrkräfte bei sehr leichten Aufgaben die Schülerleistungen tendenziell unterschätzten.

Differenzierungskomponente

Hinsichtlich der Differenzierungskomponente zeigt sich regelhaft eine Unterschätzung der Aufgabenheterogenität durch die Lehrkräfte. Die Unterschätzung ist für offene Formate in jedem Jahr signifikant ausgeprägter als für die geschlossenen: Die Einschätzung der Heterogenität der Aufgabenschwierigkeit gelingt somit bei geschlossenen Formaten besser als bei offenen. Dieser Befund entspricht nicht den als Hypothesen formulierten Erwartungen.

Schult und Lindner halten es aufgrund der geringeren tatsächlichen Varianz der Aufgabenschwierigkeiten in geschlossenen gegenüber offenen Formaten für möglich, dass dieser erwartungswidrige Befund durch die unterschiedlich ausgeprägte tatsächliche Heterogenität bedingt sein könnte, daher seien Ergebnisse zur Differenzierungskomponente mit Vorsicht zu interpretieren.

Rangordnungskomponente

Die Rangordnungskomponente variiert sowohl zwischen den Klassen innerhalb eines Jahres als auch zwischen den Jahren deutlich. Meist ist die mittlere Korrelation – wie in der Hypothese vorhergesagt – bei den offenen Formaten größer als bei den geschlossenen, nur im Jahr 2014 kehrt sich dieses Verhältnis um.

Der Autor und die Autorin resümieren, dass das Antwortformat aufgrund der Erkenntnisse aus ihrer Studie einen Faktor darstellen dürfte, der Einfluss auf die Diagnoseakkuratheit von Lehrkräften nimmt.

Diskussion und Einschätzung

Hintergrund

Zu Recht schätzen Schult und Lindner die Fähigkeit, Schülerleistungen und Aufgabenanforderungen akkurat erkennen zu können, als wesentliche Kompetenz von Lehrkräften ein und so ist es nur ein folgerichtiger Schritt, diesbezüglich nach Einflussfaktoren wie dem Antwortformat zu fragen, zumal dazu

noch keine belastbaren Befunde in der Forschung vorliegen.

Design

Der methodische Weg, um zu prüfen, ob die Verwendung offener oder geschlossener Formate einen Einfluss auf die Diagnoseakkuratheit hat, erscheint sinnvoll und erbringt einen interpretationsfähigen Befund.

Limitationen sehen die Autorin und der Autor unter anderem darin, dass nur für 3 – 5 % der jeweiligen Jahrgänge von VERA 3 die Diagnoseakkuratheit festgestellt wurde. Damit stellt sich das Problem der Repräsentativität der betroffenen Grundschullehrkräfte. Auch konnte aus Datenschutzgründen nicht ausgeschlossen werden, dass eine Lehrkraft mehrere der in die Untersuchung einbezogenen Klassen betreute. Schult und Lindner halten dieses Problem jedoch für zu vernachlässigen.

Gravierender erscheint der Autorin und dem Autor, dass keine Daten zu Eigenschaften der Lehrkräfte vorliegen, etwa zu fachlichem und fachdidaktischem Wissen und der Einstellung gegenüber offenen/geschlossenen Aufgabenformaten. Hier sehen Schult und Lindner ebenso Forschungsbedarf wie in einer vergleichenden Ausweitung der Studie auf ein Land, in dem geschlossene Antwortformate im Gegensatz zu Deutschland alltäglich sind (USA!). Dadurch könnte die Bedeutung der unterschiedlichen Erfahrungen, die Lehrer mit geschlossenen Formaten haben, abgeschätzt werden. Auch sollten Studien in unterschiedlichen Fächern und Jahrgangsstufen erfolgen, um die Übertragbarkeit der bislang erzielten Ergebnisse zu überprüfen. Abschließend mahnen die Autorin und der Autor für zukünftige Studien ein ausgewogenes Anspruchsniveau von offenen und geschlossenen Aufgabenformaten an.

Ergebnisse

Praktische Konsequenzen ihrer Untersuchung sehen Autorin und Autor darin, dass zukünftig geschlossene Formate, die bislang außerhalb von Schulleistungsstudien und Vergleichsarbeiten (PISA, IGLU, VERA etc.) keine große Rolle bei der schriftlichen Leistungserfassung im schulischen Alltag spielen, stärker für eine effiziente Lernverlaufsdiagnostik genutzt werden könnten. Offene Formate hingegen könnten eher zur Erfassung kreativ-schöpferischer Leistungen beitragen. Hierzu müsste aber in der Fortbildung von Lehrkräften eine gezielte Wissensvermittlung über diagnostische Eigenschaften und Möglichkeiten geschlossener Formate erfolgen. Auch sehen Schult und Lindner Möglichkeiten, die Selbstreflexion der Lehrkräfte bezüglich der eigenen diagnostischen Kompetenzen anzuregen und eine Auseinandersetzung mit Aufgabenformaten zu fördern, indem bei zukünftigen Vergleichsarbeiten die Urteilsakkuratheit der geschlossenen und offenen Formate jeweils aufgeschlüsselt an die Lehrkraft zurückgemeldet wird. So nachvollziehbar dieses Plädoyer von Schult und Lindner ist, so ist doch anzumerken, dass sein Inhalt nicht allzu eng mit den Ergebnissen ihrer Untersuchung zusammenhängt, geschweige denn sich zwingend aus ihnen ergibt.

Einerseits halten Schult und Lindner somit geschlossene Formate offensichtlich für noch zu wenig für effiziente Diagnosen eingesetzte Instrumente und regen daher eine stärkere Auseinandersetzung der Lehrkräfte mit diesem Format an. Andererseits weisen sie aber mehrfach darauf hin, dass bei Multiple-Choice-Tests die Ergebnisse (relativ erfolgreich) erraten werden können, was ihre Tauglichkeit als Diagnoseelement deutlich einschränkt. Vielleicht ist dies auch der Grund, warum Multiple-Choice-Aufgaben an der Schule – abseits von Vergleichsarbeiten wie z. B. VERA 3 – vergleichsweise selten zum Einsatz kommen. Ob es unter diesen Umständen wünschenswert wäre, dieses Aufgabenformat verstärkt einzusetzen, ist zumindest fraglich.

Die Autorin und der Autor weisen zu Recht darauf hin, dass eine Untersuchung des Einflusses des

Aufgabenformats für andere Schülerpopulationen als den dritten Jahrgang im Fach Deutsch anzuraten wäre. Das erscheint auch dringend geboten, denn die ausgewerteten VERA-3-Tests haben ihre Tücken, welche bei zukünftigen Studien gegebenenfalls eine vollkommene methodische Umorientierung zur Folge haben müssten: Vergleichsarbeiten wie VERA 3 kennen nur richtige oder falsche Antworten, selbst eine nur teilweise richtig beantwortete Frage wird als falsch bewertet. Nur auf dieser Basis ist Diagnoseakkuratheit so zu bestimmen, wie es in dieser Arbeit geschieht. Abseits der Vergleichsarbeiten spielen bei schriftlichen Leistungsüberprüfungen jedoch – zumindest im Deutschunterricht, aber wohl auch in anderen Fächern – reine Richtig-falsch-Entscheidungen im Verlauf der Schullaufbahn eine immer geringere Rolle. Stattdessen gehen andere Punkte in die Diagnose ein, bei denen die akkurate Bewertung wesentlich schwieriger zu standardisieren und damit formelartig zu überprüfen sein dürfte: inhaltliche Ordnung, Folgerichtigkeit der Argumentation, Angemessenheit des Sprachstils, abstrakte interpretatorische Leistung, Beherrschung des Fachvokabulars usw. Die mit diesen Punkten verbundenen Aufgabenformate beherrschen gegen Ende der Schullaufbahnen weitestgehend die schriftlichen Überprüfungen, wohingegen reine Richtig-falsch-Entscheidungen dann kaum noch gefragt sind. Zwar handelt es sich in jedem Fall um offene Formate, abgesehen von dieser Eigenschaft haben sie jedoch kaum etwas miteinander gemein, sodass die Aussagen der Studie von Schult und Lindner zur Diagnoseakkuratheit beim Einsatz offener Formate kaum übertragbar sein dürften.

Darüber hinaus bleibt eine ganze Reihe von Fragen offen, die zu klären wären, bevor weitreichende Schlussfolgerungen gezogen werden können. So ist nicht bekannt, ob die unterschiedlichen Aufgabenformate annähernd den gleichen Schwierigkeitsgrad hatten und welche Auswirkungen auf das Ergebnis der Studie ein evtl. ungleicher Schwierigkeitsgrad gehabt haben könnte. Zudem werden mögliche statistische Zusammenhänge (etwa zwischen der Streuung von Werten und daraus abzuleitenden Konsequenzen für die Rangordnungskomponente) nicht diskutiert.

Auch deuten die Autorin und der Autor an, dass es – über das Antwortformat hinaus – weitere Faktoren geben könnte, welche die Diagnoseabhängigkeit modifizieren können. So ist z. B. nichts über relevante Eigenschaften der Lehrpersonen, welche die Einschätzungen abgaben, bekannt. Deren Erfahrungen (in Bezug auf die Fähigkeit, Schülerleistungen zu prognostizieren, und im Hinblick auf die hier eingesetzten Aufgabenformate) und deren Wissen (etwa im Hinblick auf fachdidaktische Zusammenhänge) dürften aber vermutlich das Untersuchungsergebnis beeinflussen. Hier besteht dringender Forschungsbedarf. Solange hierzu – und zur gegenseitigen Beeinflussung der Faktoren – noch nichts Konkretes bekannt ist, sind die Befunde von Schult und Lindner als vorläufig zu betrachten. Das schmälert allerdings keineswegs ihren Wert, sondern die vorliegende „Pionierstudie“ bietet vielmehr eine Basis für ausgedehnte zukünftige Arbeiten.

Reflexionsfragen für die Praxis

Nachfolgende Reflexionsfragen sind ein Angebot, die Befunde der rezensierten Studie auf das eigene Handeln als Lehrkraft oder Schulleitungsmitglied zu beziehen und zu überlegen, inwiefern sich Anregungen für die eigene Handlungspraxis ergeben. Die Befunde der rezensierten Studien sind nicht immer generalisierbar, was z. B. in einer begrenzten Stichprobe begründet ist. Aber auch in diesen Fällen können die Ergebnisse interessante Hinweise liefern, um über die eigene pädagogische und schulentwicklerische Praxis zu reflektieren.

Reflexionsfragen für Lehrkräfte:

- Bin ich mir bewusst, dass ein Zusammenhang von Aufgabenformat und Diagnoseakkuratesse bestehen kann?
- Wie schätze ich meine eigene Diagnoseakkuratesse für unterschiedliche Aufgabenformate ein?
- Welche Aufgabenformate nutze ich zur Diagnose, welche schätze ich als begrenzt geeignet/ungeeignet ein?
- Halte ich Multiple-Choice-Tests für ein geeignetes Instrument der Leistungserfassung?
- Hinsichtlich der Konstruktion oder diagnostischen Nutzung welcher Aufgabenformate habe ich Informations- oder Fortbildungsbedarf?

Reflexionsfragen für Schulleitungen:

- In welchem Ausmaß lässt sich die Diagnosefähigkeit und -genauigkeit der Lehrkräfte gezielt durch Fortbildungsmaßnahmen steigern?
- Welche Rolle können dabei Fortbildungsmaßnahmen bezüglich konkreter Aufgaben-/Antwortformate spielen?

Rezensent/-in

Dr. Heinz Sander, Lehrer am Gymnasium der Stadt Kerpen – Europaschule und Privatdozent an der Universität zu Köln

Zitiervorschlag

Sander, H. (2020). Rezension zu Schult, J. & Lindner, M. A. (2018). Diagnosegenauigkeit von Deutschlehrkräften in der Grundschule: Eine Frage des Antwortformats? Zeitschrift für Pädagogische Psychologie, 32(1–2), 75–87. *Forschungsmonitor Schule, 88*. Abgerufen von https://www.forschungs-monitor-schule.de/print.php?id=101

Urheberrecht