

Wie Physiklehrkräfte Schülertexte beurteilen – Instrumententwicklung

Im Physikunterricht werden Schüler_innen nicht nur mit fachlichen, sondern auch mit hohen sprachlichen Anforderungen konfrontiert (Bach, 1984; Kulgemeyer, 2010). Von erfolgreichen Schüler_innen erwarten Physiklehrkräfte oftmals selbstverständlich, dass sie den Umgang mit bildungs- und fachsprachlichen Registern beherrschen (Feilke, 2012; Schleppegrell, 2004). Die Arbeiten von Tajmel (2010; 2011) legen nahe, dass sich diese Erwartungshaltung auch auf Leistungsbewertungen und Leistungsurteile niederschlägt. Genauer: Es lässt sich begründet vermuten, dass Physiklehrkräfte Urteile über fachliche und sprachliche Leistungen konfundieren. Wie Lehrkräfte bei der Beurteilung von Leistungsaufgaben tatsächlich vorgehen, welchen Logiken sie dabei folgen, welche Ressourcen sie aktivieren und welche Maßstäbe sie für relevant halten, ist allerdings weitgehend unbekannt (Leuders et al., 2014). In diesem Beitrag wird die Entwicklung eines Testinstruments vorgestellt, das dazu dienen soll, die eben beschriebenen Desiderate in einer Studie mit Physiklehrkräften zu explorieren.

Aufbau des Testinstruments

Zunächst soll die finale Version des Testinstruments kurz beschrieben werden, über dessen Entwicklung hier berichtet wird: Die zentrale Idee hinter dem Aufbau des Instruments ist die Erhebungssituation so authentisch wie möglich zu gestalten. Die teilnehmenden Lehrkräfte werden daher darum gebeten, einen Erwartungshorizont zu einer Leistungsaufgabe so zu erstellen, wie sie dies unter normalen Umständen auch tun würden. Die Leistungsaufgabe fordert Schüler_innen dazu auf, ein physikalisches Problem der Akustik in Form eines Textes zu erklären. Mit Hilfe ihres Erwartungshorizonts korrigieren die Lehrkräfte anschließend 4 auf sprachlicher und auf fachlicher Ebene stark unterschiedliche Schülerlösungen. Während der gesamten Erhebungssituation werden die Lehrkräfte zu lautem Denken aufgefordert. Transkripte von Audiographien bilden die Datenbasis. Die Datenanalyse ermöglicht schlussendlich Einblicke in die Denkprozesse der Lehrkräfte während der Materialbearbeitung, die durch bloße Beobachtung nicht zugänglich sind (Heine & Schramm, 2016).

Instrumententwicklung

Die Hauptaufgabe bei der Instrumententwicklung bestand, darin eine geeignete Leistungsaufgabe zu identifizieren und zugehörige Schülerlösungstexte zu generieren. Leistungsaufgabe und Lösungstexte sollten auf die teilnehmenden Lehrkräfte authentisch wirken. Die Lösungstexte sollten sich sowohl auf fachlicher als auch auf sprachlicher Ebene stark unterscheiden. Die wesentlichen Arbeitsschritte des mehrstufigen Verfahrens waren hierbei die folgenden (für eine detaillierte Darstellung siehe Feser et al., angenommen):

- Klassenarbeitsaufgaben sammeln, pilotieren und eine geeignete Aufgabe für das weitere Vorgehen wählen. Die schlussendlich verwendete Leistungsaufgabe stammt aus einer Klassenarbeit einer Lehrkraft.
- Lösungstexte von Schüler_innen zur gewählten Leistungsaufgabe erheben (N = 116, Klassenstufe 9).
- Entwicklung zweier Kategoriensysteme, durch die fachliche bzw. sprachliche Qualitätsmerkmale der erhobenen Schülerlösungstexte getrennt erfassen werden können.

- Erste Vorauswahl von Prototypen kontrastierender Lösungstexte auf Grundlage eines ersten Ratings des Gesamtmaterials, wobei 32 % doppelcodiert wurden.
- Finale Auswahl der 4 am besten geeigneten Lösungstexte (sprachlich stark/fachlich stark, sprachlich stark/fachlich schwach, sprachlich schwach/fachlich stark, sprachlich schwach/fachlich schwach) aus der Vorauswahl mit Hilfe eines zweiten Ratings durch insgesamt 6 geschulte Codierer_innen.

Kategoriensysteme zur Erfassung fachlicher und sprachlicher Qualitätsmerkmale

Wie sich aus obiger Auflistung bereits erschließen lässt, war die Ausarbeitung zweier Kategoriensysteme, mit denen die Lösungstexte der Schüler_innen codiert wurden, der mit Abstand wichtigste Arbeitsschritt der Instrumententwicklung. Erst mit Hilfe dieser Systeme war es möglich, aus dem Gesamtmaterial eine Komposition von 4 Lösungstexten zu identifizieren, die sich gesichert auf fachlicher und sprachlicher Ebene stark voneinander unterscheiden. Um diesem Zweck zu genügen, sind die Kategoriensysteme hierarchisch und graduiert aufgebaut. Das bedeutet: Beide Kategoriensysteme unterscheiden verschiedene fachliche bzw. sprachliche Qualitätsmerkmale eines Lösungstextes, wobei bei jedem Qualitätsmerkmal 2 bis 3 Ausprägungen unterschieden werden. Diesen Ausprägungen sind wiederum Punktwerte zugeordnet. Beim Rating der Lösungstexte gilt dann die Codierregel, dass jedem Lösungstext bzgl. jedem Qualitätsmerkmal genau eine Ausprägung zuzuweisen ist. Wie aus der Übersicht der Kategoriensysteme in Tabelle 1 und 2 deutlich wird, werden durch den Codiervorgang jedem Lösungstext insgesamt 4 Punktwerte bzgl. fachlicher Qualitätsmerkmale und 2 Punktwerte bzgl. sprachlicher Qualitätsmerkmale zugewiesen. Durch Aufsummieren der Punkte je Kategoriensystem erhält man schließlich einen Summenscore für die fachliche und einen Summenscore für die sprachliche Qualität je Schülerlösungstext. Für den Summenscore der fachlichen Qualität hat es sich als zielführend erwiesen, den Punktwert eines Lösungstextes des Qualitätsmerkmals „Konsistenz der Erklärung“ zusätzlich mit dem Faktor 3 zu gewichten. Durch diese Gewichtung wird gewährleistet, dass ein Lösungstext, der in den 3 anderen fachlichen Qualitätsmerkmalen volle Punktzahl erreicht, aber den physikalischen Sachverhalt nicht korrekt erklärt, höchstens 50 % der Gesamtpunktzahl erreichen kann.

Entwickelt wurden die Kategoriensysteme durch deduktiv-induktives Vorgehen (Saldaña 2015). Das Kategoriensystem bzgl. fachlicher Qualitätsmerkmale eines Lösungstexte basiert auf den Arbeiten von Kang et. al. (2014) und Braaten & Windschitl (2011) sowie der sog. SOLO-Taxonomie von Biggs & Collis (1982). Die Grundlage für das Kategoriensystem für die sprachliche Qualität eines Lösungstextes bildete das Raster zur Beobachtung und Analyse bildungssprachlicher Fähigkeiten der FörMig-Initiative (Roth et al., 2011; Lengyel et al., 2009), wobei hier u. A. Aspekte aus den Arbeiten von Tajmel (2011) und Koch & Oesterreicher (1985) ergänzt wurden. Im Rahmen der Instrumententwicklung konnte die Intercoderreliabilität der einzelnen Kategorien insgesamt zwei Mal überprüft werden, wobei sich beide Male gute bis sehr gute Ergebnisse ergaben. Im Rahmen der Doppelcodierung des ersten Ratings lag diese zwischen $\alpha = 0.88$ und $\alpha = 0.92$, wobei hier die Kategoriensysteme z. T. noch missverständlich formuliert waren und deswegen unterschiedliche Codierungen intensiv diskutiert werden mussten. Durch Verbesserung der Kategoriensysteme konnte im zweiten Rating auf eine Diskussion verzichtet werden und es ergab sich eine Intercoderreliabilität zwischen $W = 0.72$ und $W = 0.94$.

Resümee

Es wurde ein Testinstrument entwickelt, das zu explorieren ermöglicht, wie Physiklehrkräfte Schülertexte beurteilen. Ferner konnten zwei reliable Kategoriensysteme entwickelt werden, die es ermöglichen, schriftliche Schülerlösungstexte einer physikalischen Leistungsaufgabe sowohl auf fachlicher, als auch auf sprachlicher Ebene unterscheiden zu können.

Qualitätsmerkmal	Ausprägung 1 (0 Punkte)	Ausprägung 2 (1 Punkt)	Ausprägung 3 (2 Punkte)
Rahmenbau der Erklärung	Der Text ist eine bruchstückhafte Wiedergabe des Aufgabestamms oder eine bloße Paraphrasierung von Merksätzen. Er ist nicht/kaum empirie- oder theoriegeleitet.	---	Argumentation durch empirische Daten, Erfahrungen oder Anwendung physikalischen Sachwissens. Der generalisierende Charakter des Textes ist dominant.
Rolle von Evidenzbezeugen	Der Text enthält keine Evidenzmittel zur Stützung von Aussagen.	Die verwendeten Evidenzmittel sind entweder nicht angemessen oder hinlänglich.	Die verwendeten Evidenzmittel sind sowohl angemessen als auch hinlänglich.
Tiefe der Erklärung	Erklärung erfolgt auf der „Was“-Ebene.	Erklärung erfolgt auf der „Wie“-Ebene.	Erklärung erfolgt auf der „Warum“-Ebene.
Konsistenz der Erklärung	Sachverhalte, die nicht im Fokus der Aufgabe stehen, werden erklärt oder fachlich falsche Erklärungen werden gegeben.	Die Erklärung ist teilweise korrekt. Teile der Erklärung sind übergeneralisiert, irrelevant oder unvollständig.	Der Sachverhalt, der im Fokus der Aufgabe steht, wird allumfassend und fachlich richtig erklärt.

Tab. 1: Kategoriensystem zur Unterscheidung fachlicher Qualitätsmerkmale eines Lösungstextes (Biggs & Collis, 1982; Braaten & Windschitl, 2011; Kang et al., 2014).

Qualitätsmerkmal	Ausprägung 1 (0 Punkte)	Ausprägung 2 (1 Punkt)	Ausprägung 3 (2 Punkte)
Lexik/Semantik	Ein Alltagswortschatz mit dominant mündlichem Charakter wird verwendet.	Ein Alltagswortschatz wird verwendet, der sich deutlich von einer mündlichen Ausdrucksweise absetzt.	Ein unterrichtlicher Fachwortschatz wird verwendet.
Syntax/Stilistik	Es werden nur einfache/unvollständige Sätze oder einfach Satzreihen gebildet. Stilistisch entspricht der Text tendenziell einer Sprache der Nähe.	Unterschiedliche satzverbindende Elemente werden eingesetzt. Sprachliche Verdichtung findet nicht/kaum statt. Stilistisch entspricht der Text tendenziell einer Sprache der Nähe.	Die Sätze sind sprachlich verdichtet und unterschiedlich verbunden. Stilistisch entspricht der Text tendenziell einer Sprache der Distanz.

Tab. 2: Kategoriensystem zur Unterscheidung sprachlicher Qualitätsmerkmale eines Lösungstextes (Roth et al., 2011; Koch & Oesterreicher, 1985; Lengyel et al., 2009; Tajmel, 2011).

Literatur

- Bach, S. (1984). Systematische und empirische Untersuchung über das Verhältnis von Umgangssprache und Fachsprache im gymnasialen Physikunterricht. Dissertation. Universität Hamburg.
- Biggs, J.B., & Collis, K.F. (1982). Evaluating the Quality of Learning. The SOLO Taxonomy (Structure of the Observed Learning Outcomes). Educational Psychology Series, Academic Press.
- Braaten, M., & Windschitl, M. (2011). Working Toward a Stronger Conceptualization of Scientific Explanation for Science Education. *Science Education*, 95 (4), 639-669.
- Feilke, H. (2012). Schulsprache - Wie Schule Sprache macht. In S. Günther, W. Imo, & D. Meer, J.G. Schneider (Eds.), *Kommunikation und Öffentlichkeit. Sprachwissenschaftliche Potenziale zwischen Empirie und Norm* (pp. 151-175) De Gruyter.
- Feser, M.S., Höttecke, D., & Ehmke, T. (angenommen). Testitems zur qualitativen Untersuchung der Ressourcen von Physiklehrkräften beim Bewerten schriftlicher Schülerleistungen. *PhyDid B - Didaktik der Physik - Beiträge zur DPG-Frühjahrstagung*, o. V. (o. N.), o. S..
- Heine, L., & Schramm, K. (2016). Introspektion. In K. Schramm, K. Schramm, & F. Klippel, M.K. Legutke (Eds.), *Forschungsmethoden in der Fremdsprachendidaktik. Ein Handbuch* (pp. 173-181) Narr Francke Attempto Verlag.
- Kang, H., Thompson, J., & Windschitl, M. (2014). Creating Opportunities for Students to Show What They Know: The Role of Scaffolding in Assessment Tasks. *Science Education*, 98 (4), 674-704.
- Koch, P., & Oesterreicher, W. (1985). Sprache der Nähe - Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In G. Folke, D. Jacob, & A. Kablitz, B. König, K. Joachim, D. Nelting, C. Schmitt, M. Selig, S. Zepp (Eds.), *Romanistisches Jahrbuch* (pp. 15-43) [Romanistisches Jahrbuch, Vol. 36].
- Kulgemeyer, C. (2010). *Physikalische Kommunikationskompetenz. Modellierung und Diagnostik*. Logos Verlag Berlin.
- Lengyel, D., Reich, H.H., & Roth, H.J., Heintze, A., Scheinhardt-Stettner, H. (2009). Prozessbegleitende Diagnose zur Schreibentwicklung: Beobachtung schriftlicher Sprachhandlungen in der Sekundarstufe I. In D. Lengyel, H.H. Reich, & H.J. Roth, M. Döll (Eds.), *Von der Sprachdiagnose zur Sprachförderung* (pp. 129-138) Waxmann.
- Leuders, T., Philipp, K., & Leuders, J. (2014). Fachbezogene diagnostische Kompetenzen - Forschungsstand und Forschungsdesiderata. In J. Roth, & J. Ames (Eds.), *Beiträge zum Mathematikunterricht 2014* (pp. 731-734) Verlag für wissenschaftliche Texte und Medien Münster.
- Roth, H.J., Lengyel, D., & Reich, H.H. (2011). *Sprachhandlungen Erklären, Berichten und Argumentieren. Manual zu den Auswertungsrastern. Entwurfsfassung*. FÖRMIG AG SEK I.
- Saldaña, J. (2015). *The Coding Manual for Qualitative Researchers*. SAGE Publications Inc..
- Schleppegrell, M.J. (2004). *The Language of Schooling. A Functional Linguistics Perspective*. Routledge.
- Tajmel, T. (2010). DaZ-Förderung im naturwissenschaftlichen Fachunterricht. In B. Ahrenholz (Ed.), *Fachunterricht und Deutsch als Zweitsprache* (pp. 167-184) 1 ed. Narr Francke Attempto Verlag.
- Tajmel, T. (2011). Wortschatzarbeit im mathematisch-naturwissenschaftlichen Unterricht. *ide*, 35 (1), 83-93.