Markus Sebastian Feser Universität Hamburg
Dietmar Höttecke

## Physiklehrkräfte beurteilen Schülertexte – Eine Explorationsstudie

Die Beurteilung der Leistung von Schüler_innen ist ein bedeutsamer Aspekt der täglichen Berufspraxis von Lehrkräften. Bisherige Untersuchungen aus einer eher allgemeindidaktischen Perspektive haben gezeigt, dass Sekundarstufenlehrkräfte für die Leistungsbeurteilung von Schüler_innen tendenziell vor allem auf Klassenarbeiten zurückgreifen und auf diese als Informationsquelle vertrauen (vgl. Marso & Pigge, 1993). Des Weiteren deutet die bisherige Forschung darauf hin, dass sich Lehrkräfte Strategien und Kriterien, die dazu dienen, die Leistung von Schüler_innen im Rahmen einer Klassenarbeit zu beurteilen, vor allem in ihrem Berufsalltag, jedoch kaum im Rahmen von theoretischer Aus- und Weiterbildung aneignen (vgl. Terhart, 2000). Generell ist jedoch festzustellen, dass alltägliche Leistungsbeurteilung durch Lehrkräfte einen Gegenstand dargestellt, mit dem sich die erziehungswissenschaftliche und insbesondere physikdidaktische Forschung bislang nur sehr wenig auseinandergesetzt hat (z. B. Stiggins, 1991).
Aus der fachdidaktischen Forschung ist bekannt, dass der Physikunterricht neben fachlichen auch hohe sprachliche Anforderungen an Schüler_innen stellt (z. B. Riebling 2013; Rincke, 2007). Nach Wellington & Osborne (2001) ist daher «[e]very science lesson [...] a language lesson» (S. 2), weswegen Sprache eine zentrale Zugangsbarriere für naturwissenschaftliche Bildung darstellt (vgl. ebd.; Tajmel, 2017). Es ist daher naheliegend davon auszugehen, dass die Beurteilung schriftlicher Schülerleistungen durch Physiklehrkräfte hiervon nicht unberührt bleibt. An Fallbeispielen konnte Tajmel (2010) bereits darlegen, dass Physiklehrkräfte bei fachlicher Leistungsbeurteilung hohe Erwartungen an die sprachliche Form stellen und beim Fällen fachlicher Leistungsurteile mitbewerten. Wir vermuten, dass sich dies auch auf die Beurteilung von Klassenarbeiten niederschlägt. Darüber, wie Physiklehrkräfte bei der Beurteilung einer Klassenarbeit vorgehen und ob hierbei fachliche und sprachliche Urteile konfundieren, liegt allerdings bis dato keine belastbare empirische Evidenz vor. Im Projekt „Fachliche und sprachliche Urteilskriterien von Physiklehrkräften“ explorieren wir daher die Genese von Urteilen über schriftliche Schülerlösungen aus einer Leistungssituation. Zusammengefasst interessieren uns die folgenden beiden Forschungsfragen in diesem Projekt:

1. Welche Ressourcen werden von Physiklehrkräften zur fachlichen und sprachlichen Beurteilung schriftlicher Leistungsaufgaben eingesetzt?
2. Inwieweit findet beim Beurteilen von Schülerleistungen eine Konfundierung fachlicher und sprachlicher Leistungsurteile statt?

### Methodisches Vorgehen

Um diese beiden Fragen explorieren zu können, haben wir im Rahmen einer Vorstudie ein geeignetes Erhebungsinstrument für Physiklehrkräfte entwickelt. Dessen finale Version soll nun zunächst kurz beschrieben werden (für eine detaillierte Darstellung der Entwicklungsarbeit siehe Feser et al. (2016), sowie Feser & Höttecke (2017a)): Die zentrale Idee hinter dem Aufbau des Instruments ist die Erhebungssituation so authentisch wie möglich zu gestalten. Die teilnehmenden Lehrkräfte werden daher im ersten Teil der Erhebung 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 Phänomen der Akustik in Form eines Textes zu erklären. Mit Hilfe ihres Erwartungshorizonts korrigieren die Lehrkräfte anschließen vier auf sprachlicher und auf fachlicher Ebene stark unterschiedliche Schülerlösungen. Geeignete Schülerlösungen sind in einer Vorstudie generiert worden (vgl. Feser & Höttecke, 2017b). In einem sich anschließenden Postinterview wird die fachliche und sprachliche Qualität der zuvor von den Lehrkräften korrigierten Schülerantworten noch einmal beleuchtet. Dazu werden den Lehrkräften die Schülerantworten in Paaren vorgelegt. Bei vier verschiedenen Schülerantworten gibt es insgesamt sechs unterschiedliche Paare. Jedes dieser sechs Paare wird den Lehrkräften zweimal vorgelegt. Beim ersten (zweiten) Mal erhalten sie die Instruktion:

> «Beurteilen Sie, ob eine der beiden Antworten fachlich (sprachlich) besser ist, oder ob sie fachlich gleich gut sind. Ob evtl. eine der beiden Antworten sprachlich (fachlich) besser ist, soll hierbei komplett unberücksichtigt bleiben. Bitte begründen Sie Ihre Entscheidung.»

Audiographien der Think-Aloud-Aufgabe und des Postinterviews bilden die Datenbasis der Hauptstudie dieses Projekts. Die Datenanalyse ermöglicht schlussendlich Einblicke in die Denkprozesse der Lehrkräfte während der Materialbearbeitung, die durch bloße Beobachtung nicht zugänglich sind (Heine & Schramm, 2016). Vor der eigentlichen Korrekturarbeit findet zudem ein intensives Training der Think-Aloud-Methode statt, um die Validität der erhobenen Daten sicher zu stellen (Heine & Schramm, 2007; van Someren et al. 1994).

Die Erhebung der Hauptstudie, in der das eben beschriebene Instrument eingesetzt wurde, fand von April bis September 2016 statt. Insgesamt wurde dabei ein heterogenes[1] Gelegenheitssample von N=21 Hamburger Physiklehrkräfte gewonnen. Parallel zur Erhebungsphase wurden von den Audiographien der Think-Aloud-Aufgabe und des Postinterviews manualgeleitete Detailtranskripte angefertigt (vgl. Fuß & Karbach, 2014). Anschließend wurde damit begonnen, die Laut-Denk-Protokolle und die Transkripte der Postinterviews zunächst getrennt voneinander auszuwerten. Die Entscheidung, den erhobenen Datensatz zunächst in zwei Teildatensätze zu zerlegen, erfolgte, um dem Umstand gerecht zu werden, dass die Postinterviews Einblicke in die «reflective perspective» von Physiklehrkräften auf die Beurteilung schriftlicher Schülerleistungen gewähren, wohin gehen die Laut-Denk-Protokolle die «in-action perspective» beleuchten (vgl. Lindmeier, 2011). Die Analysen der Datensätze erfolgte ferner unter zu Hilfenahme von sowohl qualitativen als auch quantitativen Methoden: Die Verbaldaten wurden mit Hilfe verschiedener Techniken der qualitativen Inhaltsanalyse ausgewertet (vgl. Mayring, 2015) und sofern dies möglich bzw. zulässig war, wurden die so gewonnen qualitativen Befunde quantifiziert (vgl. Kuckartz, 2014) und mit Hilfe non-parametrischer statistischer Methoden weiter analysiert (vgl. Siegel, 1976). Wir berichten hier ausgewählte Befunde, die wir aus der Analyse der Postinterviews gewonnen haben:

[1] Es wurden Physiklehrkräfte mit unterschiedlich langer Berufserfahrung (2.5 bis 37 Jahre), mit verschiedenen Zweitfächern (insbesondere sprachlichen Fächern) und die an Gymnasien und/oder Stadtteilschulen unterrichten befragt.

**Befunde aus der Analyse der Postinterviews**

In einem ersten Analyseschritt wurden die Postinterview-Transkripte einer inhaltlich strukturierenden Inhaltsanalyse (Mayring, 2015) unterzogen. Dieses Verfahren ermöglichte es uns, die von den befragten Lehrkräften für die Beurteilungen der Schülertexte verwandten Beurteilungskriterien induktiv zu identifizieren. Dabei zeigte sich, dass die befragten Lehrkräfte für die Paarvergleiche bzgl. der fachlichen Qualität zweier Schülertexte insgesamt 13 verschiedene Beurteilungskriterien und für die Paarvergleiche bzgl. der sprachlichen Qualität zweier Schülertexte insgesamt 20 verschiedene Beurteilungskriterien eingesetzt haben[2]. Insbesondere wurde allerdings sichtbar, dass die befragten Lehrkräfte 7 Kriterien in beiden Paarvergleichsaufgaben eingesetzt haben. Wortwörtlich waren diese, «die Entsprechung meiner persönlichen Erwartungen», «die Quantität von Fachwörtern», «die Qualität der (Fach-)Sprache», «der/die Verdichtungsgrad/Präzision des Textes», «die Strukturiertheit/Gliederung des Textes», «die Differenziertheit/Komplexität des Textes» und «das Vorhandensein von Redundanz». Aus unserer Sicht besonders hervorzuheben sind dabei die hier an zweiter und dritter Stelle aufgeführten Kriterien. Bei beiden handelt es sich eindeutig um Beurteilungskriterien, die die sprachliche Realisierung eines Schülertextes betreffen, was umso bemerkenswerter ist, als dass diese von den befragten Lehrkräften (trotz explizit andere Aufforderung!) dazu verwandt wurden, Unterschiede zwischen zwei Schülertexten bzgl. ihrer fachlichen Qualität ggf. zu begründen. Anders ausgedrückt zeigt sich hier also ein sehr deutlicher empirischer Hinweis einer Konfundierung der fachlichen und sprachlichen Leistungsbeurteilung, wie Eingangs von uns vermutet.

Wie ausgeprägt diese Konfundierung ist, darüber lässt sich mit Hilfe des von uns gewählte qualitative Vorgehen jedoch keine Aussage treffen. Aus diesem Grund wurden die in den Postinterviews gewonnenen Daten einer zusätzlichen quantitativen Analyse unterzogen. Hierzu wurden die von jeder Lehrkraft vorgenommenen Beurteilungen durch paarweisen Vergleich in zwei Rangreihen der 4 Schülertexte bzgl. ihrer fachlichen bzw. ihrer sprachlichen Qualität quantifiziert und mit Hilfe dieser Rangreihen anschließend das von Togerson (1956) und Ludwig (1962) vorgeschlagene Rangkorrelationsmaß $\tau^*$ bestimmt. Vereinfacht ausgedrückt liefert dieses non-parametrische Korrelationsmaß eine Schätzung der „mittleren“ Stärke des monotonen Zusammenhangs zwischen der fachlichen und sprachlichen Leistungsbeurteilung der befragten Lehrkräfte im Rahmen der Postinterviews. Unsere Berechnung liefert dabei einen Wert von $\tau^*=.42$, der zum Signifikanzniveau $\alpha=.01$ verschieden von 0 ist. Hierbei ist zu beachten, dass die 4 Schülertexte im Rahmen der Vorstudie derart ausgewählt wurden, dass ein Wert von $\tau^*\approx 0$ theoretisch zu erwarten gewesen wäre, der sich hätte zeigen müssen, wenn die Lehrkräfte fachliche und sprachliche Urteile nicht konfundieren.

**Resümee**

Wie aus dem vorangegangen Abschnitt deutlich wurde, liefert bereits die Analyse der im Rahmen der Postinterviews erhobenen Daten einen reichhaltigen Einblick in die bislang kaum erforschte Genese von Leistungsurteilen von Physiklehrkräften. Zusätzlich hervorzuheben ist, dass die aufgedeckten empirischen Hinweise auf eine Konfundierung von fachlicher und sprachlicher Leistungsbeurteilung als ein Indiz auf eine mangelnde «kritisch-reflexive Sprachbewusstheit» (vgl. Tajmel, 2017) der von uns befragten Physiklehrkraft aufgefasst werden kann und damit deren professionelles Handeln im Kontext von schulischer Leistungsbeurteilung zumindest fragwürdig erscheint.

[2] Eine vollständige Auflistung dieser 13 bzw. 20 Kriterien findet sich bei Feser & Höttecke (2017c).

## Literatur


Feser, M.S., Höttecke, D., & Ehmke, T. (2016). 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..

Feser, M. & Höttecke, D. (2017a). Wie Physiklehrkräfte Schülertexte beurteilen – Instrumententwicklung. In Chr. Maurer (Hrsg.), Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis (S. 123-126). Gesellschaft für Didaktik der Chemie und Physik Jahrestagung in Zürich 2016.

Feser, M.S. & Höttecke, D. (2017b). Klassenarbeiten kriteriengeleitet korrigieren – Wie beurteile ich eine Schülererklärung?. Unterricht Physik, 158, S. 15-18.

Feser, M.S. & Höttecke, D. (2017c). How physics teachers assess students' texts in teacher-made tests. Paper presented at the ESERA conference 21st-25th August 2017 in Dublin, http://keynote.conference-services.net/resources/444/5233/pdf/ESERA2017_0099_paper.pdf (14.09.2017).

Fuß, S., & Karbach, U. (2014). *Grundlagen der Transkription. Eine praktische Einführung*.Verlag Barbara Budrich.

Heine, L., & Schramm, K. (2007). Lautes Denken in der Fremdsprachenforschung: Eine Handreichung für die empirische Praxis. In H.J. Vollmer (Ed.), *Synergieeffekte in der Fremdsprachenforschung. Empirische Zugänge, Probleme, Ergebnisse* (pp. 167-206)Europäischer Verlag der Wissenschaften.

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.

Kuckartz, U. (2014). *Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren*.Springer VS.

Lindmeier, A. (2011). *Modeling and Measuring Knowledge and Competencies of Teachers. A Threefold Domain-specific Structure Model for Mathematics*. Münster: Waxmann.

Ludwig, O. (1962). Über Kombination von Rangkorrelationskoeffizienten aus unabhängigen Meßreihen. *Biometrical Journal*, 4 (1), 40-50.

Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken*.Beltz Verlag.

Marso, R.N., & Pigge, F.L. (1993). Teachers' Testing Knowledge, Skills, and Practices. In S.L. Wise (Ed.), *Teacher Training in Measurement and Assessment Skills* (pp. 129-185), Lincoln (NE): Buros Institute of Mental Measurements.

Riebling, L. (2013). *Sprachbildung im naturwissenschafltichen Unterricht. Eine Studie im Kontext migrationsbedingter sprachlicher Heterogenität*.Waxmann.

Rincke, K. (2007). *Sprachentwicklung und Fachlernen im Mechanikunterricht. Sprache und Kommunikation bei der Einführung in den Kraftbegriff*.Logos Verlag Berlin.

Siegel, S. (1976). *Nichtparametrische statistische Methoden*.Fachbuchhandlung für Psychologie Verlagsabteilung.

Stiggins, R.J. (1991). Assessment Literacy. *Phi Delta Kappan*, 72 (7), 534-539.

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. (2017). *Naturwissenschaftliche Bildung in der Migrationsgesellschaft. Grundzüge einer Reflexiven Physikdidaktik und kritisch-sprachbewussten Praxis*. Wiesbaden: Springer VS.

Terhart, E. (2000). Schüler beurteilen - Zensuren geben. Wie Lehrerinnen und Lehrer mit einem leidigen, aber unausweichlichen Element ihres Berufsalltags umgehen. In S.I. Beutel, & W. Vollstädt (Eds.), *Leistung ermitteln und bewerten* (pp. 39-50), Hamburg: Bergmann + Helbig.

Togerson, W.S. (1956). A non-parametric test of correlation using rank orders within subgroups. *Psychometrika*, 21 (2), 145-152.

van Someren, M.W., Barnard, Y.F., & Sandberg, J.A.C. (1994). *The Think Aloud Method. A pracitcal guide to modelling cognitive processes*.Academic Press.

Wellington, J., & Osborne, J. (2001). *Language and Literacy in Science Education*.Open University Press.