Testitems zur qualitativen Untersuchung der Ressourcen von Physiklehrkräften beim Bewerten schriftlicher Schülerleistungen in Physik

Markus Sebastian Feser*, Dietmar Höttecke*, Timo Ehmke[+]

*Universität Hamburg, Arbeitsgruppe Physikdidaktik [+]Leuphana Universität Lüneburg, Institut für Bildungswissenschaft
markus.sebastian.feser@uni-hamburg.de, dietmar.hoettecke@uni-hamburg.de, timo.ehmke@uni.leuphana.de

Kurzfassung

Die Diagnostik schriftlicher Schülerleistungen ist ein wichtiger Teil der täglichen Arbeit von Physiklehrkräften. Wir hegen die begründete Vermutung, dass Physiklehrkräfte dabei ihre Urteile über fachliche und sprachliche Leistungen konfundieren. Wir wissen nicht, ob und ggf. wie sich die Bewertungen über fachliche und sprachliche Schülerleistungen beeinflussen. Zudem liegt keine belastbare empirische Evidenz darüber vor, auf welche Ressourcen Physiklehrkräfte beim Bewerten schriftlicher Schülerleistungen zurückgreifen. Diese Fragen stellen wir im Projekt „Fachliche und sprachliche Urteilskriterien von Physiklehrkräften“. In einer Vorstudie wurden zunächst 128 Schüler_innen Hamburger Gymnasien und Stadtteilschulen gebeten, eine Leistungsaufgabe aus der Akustik zu bearbeiten. Die Schülertexte wurden anschließend in ein Koordinatensystem eingeordnet, welches die Modell-Dimensionen fachliche und sprachliche Qualität aufspannt. Die Einordnung wurde über ein Codierverfahren und ein Expertenrating abgesichert. Hierdurch konnten vier kontrastierende Schülertexte identifiziert werden, die nun in einer Hauptstudie mit Physiklehrkräften als Think-Aloud-Aufgaben dienen. Dieser Artikel berichtet über das Design, die Instrumentenentwicklung und Ergebnisse der Vorstudie.

1. Problemaufriss und Forschungsstand

Im Physikunterricht werden neben fachlichen Anforderungen implizit auch hohe sprachliche Anforderungen an Schüler_innen gestellt [2][17][22]. Die Linguistik spricht allgemein von normativen Sprachregistern, deren adäquate Beherrschung von erfolgreichen Schüler_innen erwartet wird [7][24]. Dass die Diagnostik schriftlicher Schülerleistungen, die ein wichtiger Teil der täglichen Arbeit von Physiklehrkräften ist, hiervon nicht unberührt bleibt, lässt sich anhand folgender Äußerung einer Physiklehrkraft, die im Rahmen eines Interviews im April 2016 aufgenommen wurde, veranschaulichen:

> **REI1848:** *[...] [Ich] stelle fest, auch hier wieder, dass es nicht so ganz einfach ist [...] sozusagen mit dem gleichen Bewertungslevel oder Bewertungsschlüssel äh Texte zu korrigieren bzw. Texte zu bewerten. Weil's eben viele Kriterien gibt. Also sprachliche sind dann vielleicht eher nachgeordnet, wenn's um das Fach geht. Aber eben auch Benutzung von Fachbegriffen. Ich ganz persönlich finde, das ist jetzt ein Statement keine Frage, ich gehöre schon noch zu denjenigen die eigentlich von Schülern erwarten, dass sie zumindestens formal korrekte Sätze formulieren und nicht darum was sie hinklieren.*

Zu Beginn macht die Lehrkraft deutlich, dass sie die Diagnostik schriftlicher Schülerleistungen aufgrund der Vielzahl möglicher Urteilskriterien als nicht triviale Anforderung empfindet. Anderseits präzisiert die Lehrkraft am Ende des Interviewauszugs die Erwartungshaltung ein Schülertext solle einem bestimmten normativen Sprachregister genügen. Insbesondere durch die Aussage, dass ein Schülertext *„zuminde[st] [aus] formal korrekten Sätzen*“ bestehen sollte, macht dies die Lehrkraft sehr deutlich. Ferner deckt sich das „*Statement*“ der Lehrkraft mit der von Tajmel geäußerten These, dass Physiklehrkräfte bei fachlicher Leistungsbeurteilung hohe Ansprüche an die sprachliche Form stellen und des Weiteren diese stets die sprachliche Leistungen von Schüler_innen mitbewerten [29].

Neben der Frage nach den Ressourcen zur Urteilsgenese lässt sich, unter anderem auf Grundlage der Darstellungen von Tajmel, die begründete Vermutung äußern, dass Physiklehrkräfte während der Genese der Leistungsbewertung ihre Urteile über fachliche und sprachliche Leistungen konfundieren [29][30]. Der oben zitierte Interviewausschnitt liefert hierzu ebenfalls Hinweise: Die Lehrkraft spricht im dritten Satz davon, dass „*sprachliche [Kriterien] [...] eher nachgeordnet [sind], wenn's um das Fach geht“*. Unmittelbar darauf folgend betont die Lehrkraft allerdings die Erwartung, dass Schüler_innen „*nicht [...][irgend-etwas] hinklieren*“. Das Dialektwort „*hinklieren*“ ist hier im Sinne von *„nachlässig“* oder *„unsauber schreiben“* negativ konnotiert.

Hierdurch korrigiert die Lehrkraft ihre zuvor gemachte Setzung, sprachliche Kriterien generell geringer als fachliche Kriterien zu gewichten. Ihre Bewertungslogik scheint viel mehr darin zu bestehen, fachliche Qualitätsmerkmale mehr oder weniger zu relativieren, wenn ein Schülertext zu viele Mängel bzgl. einer sprachlichen Norm aufweist, da die Beherrschung dieser Norm von der Lehrkraft als selbstverständlich vorausgesetzt wird.

Wie eine solche Relativierung von fachlichen Qualitätsmerkmalen eines Schülertextes im Fall dieser Lehrkraft aussieht, darüber lässt sich an dieser Stelle nur mutmaßen. Problematisch bzgl. einer adäquaten Diagnose könnte jedoch sein, wenn selbst fachlich richtige oder zumindest anschlussfähige Denkfiguren von Schüler_innen eine solche Relativierung erfahren.

Da dieser Interviewauszug allerdings losgelöst von einer tatsächlichen Urteilssituation ist, lassen sich an dieser Stelle keine sicheren Schlüsse darüber ziehen, wie diese Lehrkraft tatsächlich vorgeht, wenn sie Schülerleistung beurteilt, welcher Logik sie dabei folgt und welche Maßstäbe sie anwendet. Dass solche Ressourcen zur Urteilsgenese von Fachlehrkräften ein aktuelles Forschungsdesiderat darstellen, haben Leuders et al. am Beispiel der mathematikdidaktischen Forschung feststellen können [19]. Ferner lässt sich dieses Desiderat aus der Metaanalyse zum Konstrukt der diagnostischen Kompetenz von Lehrkräften von Südkamp et al. ableiten [28].

2. Das Projekt „Fachliche und sprachliche Urteilskriterien von Physiklehrkräften“

Die in Abschnitt 1 geäußerten Vermutungen lassen sich zu den folgenden zwei Forschungsfragen zusammenfassen:

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?

Diese Fragen stellen wir im Projekt „Fachliche und sprachliche Urteilskriterien von Physiklehrkräften“ in Zusammenarbeit mit der universitätsübergreifenden Arbeitsgruppe Fach und Sprache (www.fach-und-sprache.de). Zur Untersuchung dieser Vermutungen haben wir im Rahmen einer Vorstudie ein Testinstrument für Physiklehrkräfte entwickelt. Das Instrument besteht aus Think-Aloud-Aufgaben und einem Postinterview (vgl. Abb. 1). In den Think-Aloud-Aufgaben werden Lehrkräfte gebeten zu einer Leistungsaufgabe aus der Akustik einen Erwartungshorizont ihren Gewohnheiten entsprechend zu erstellen. Anschließend urteilen die Lehrkräfte mit Hilfe ihres Erwartungshorizonts über vier kontrastierende Schülerlösungen.

Abb.1: Ablaufplan der Erhebungssituation des Think-Aloud-Instruments

Vor der eigentlichen Korrekturarbeit findet zudem ein intensives Training der Think-Aloud-Methode statt, um die Validität der erhobenen Daten sicher zu stellen [11][32]. In dieser Trainingsphase betrachten die Lehrkräfte zunächst gemeinsam mit der Versuchsleitung ein Lernvideo[1] zur Think-Aloud-Methode. Abschließend üben die Lehrkräfte das laute Denken anhand eines einfachen Beispiels.

3. Schritte zur Entwicklung eines Testinstruments

Wie in Abschnitt 2 bereits erwähnt, wurden die Think-Aloud-Aufgaben als Testinstrument für Physiklehrkräfte im Rahmen dieses Projekts erst entwickelt. Für deren Entwicklungsarbeit waren zwei Leitideen von zentraler Bedeutung: Zum einen sollten die zu entwickelnden Aufgaben einen unmittelbaren Vergleich verschiedener Physiklehrkräfte ermöglichen und zum anderen sollten die Aufgaben einer realen Korrigiersituation möglichst ähnlich sein. Um dies zu gewährleisten, wurde das Erhebungsinstrument gemäß Abbildung 2 mehrstufig entwickelt.

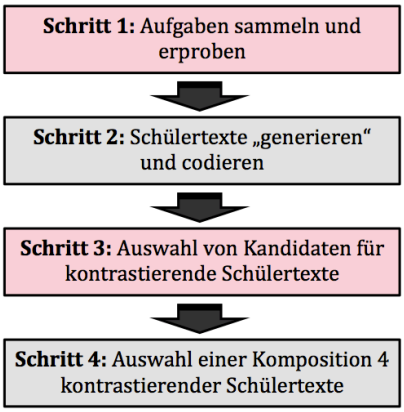

Abb.2: Schritte der Instrumentenentwicklung

Die einzelnen Entwicklungsschritte werden in den folgenden Abschnitten der Reihe nach dargestellt.

3.1. Aufgaben sammeln und erproben

Im ersten Schritt wurden verschiedene schriftliche Leistungsaufgaben gesammelt und mit dem Ziel erprobt, eine für das Testinstrument geeignete Aufgabe zu finden. Das gesetzte ökologische Validitätskriterium, ein Testinstrument zu entwickeln, das einer realen Korrigiersituation möglichst ähnlich ist,

[1]Dieses Lernvideo wurde uns von der Didaktik der Physik der Universität Potsdam für dieses Projekt zur Verfügung gestellt. Hierfür möchten wir uns herzlich bedanken.

sollte hierbei besonders berücksichtigt werden. Deshalb wurden, in Anlehnung an Thonhauser, mehrere Physiklehrkräfte gebeten, Aufgaben einzureichen, welche sie tatsächlich als Leistungsaufgaben in Klassenarbeiten eingesetzt haben [31]. Zusätzlich wurden publizierte Aufgabensammlungen von Physiklehrkräften nach geeigneten Aufgaben gesichtet (z. B. [27]).

Da im Fokus des Projekts die Diagnostik schriftlicher Schülerleistungen im Fach Physik steht, wurden folgende Auswahlkriterien für Leistungsaufgaben festgelegt:

1. Die Aufgaben fordern Schüler_innen dazu auf, einen physikalischen Sachverhalt zu erklären.
2. Die Aufgaben fordern von Schüler_innen schriftliche Lösungen mit hohem Textanteil.

Der Aufgabenpool konnte so auf 5 Leistungsaufgaben reduziert werden. Diese 5 Aufgaben wurden anschließend am Ende des Schuljahres 2014/2015 an insgesamt 23 Schüler_innen der 8. Jahrgangsstufe und 45 Schüler_innen der 9. Jahrgangsstufe verschiedener Hamburger Gymnasien und Stadtteilschulen pilotiert.

Auf Grundlage einer ersten Sichtung der Aufgabenbearbeitung wurde letztendlich die Aufgabe „Weltraumspaziergang“[2] ausgewählt (siehe Abb. 3). Diese Aufgabe entspricht den eben genannten Auswahlkriterien sowie ferner den aktuellen Mindeststandards der Hamburger Bildungspläne bzgl. inhaltlicher Mindestkompetenzen von Schüler_innen am Ende der 8. Jahrgangsstufe [9][10]. Des Weiteren zeigte sich in der Piloterhebung, dass Schüler_innen sowohl fachlich als auch sprachlich stark unterschiedliche Texte zu dieser Aufgabe produzieren und dass diese Aufgabe nur in Ausnahmefällen nicht bearbeitet wurde.

Weltraumspaziergang
Bei einem Weltraumspaziergang reißt zwischen zwei Astronauten die Funkverbindung ab. Obwohl der eine Astronaut aus Leibeskräften schreit, hört ihn sein Kamerad nicht. Der ältere Astronaut hält seinen in Panik geratenen Kollegen fest und presst seinen Helm an den des Kollegen. Plötzlich kann der jüngere den älteren leise hören. Erkläre beide Phänomene genau!

Abb.3: Aufgabe „Weltraumspaziergang“

3.2. Schülertexte „generieren“ und codieren

Ziel dieses Entwicklungsschritts war es möglichst unterschiedliche Schülertexte zur Aufgabe „Weltraumspaziergang“ zu erhalten, sowie ein Kategoriensystem zu entwickeln, dass es ermöglicht, auf sowohl sprachlicher als auch fachlicher Ebene stark kontrastierende Schülertexte zu identifizieren. Da davon auszugehen war, dass Kontrastfälle eher selten vorzufinden sind, wurde die Aufgabe zu Beginn des Schuljahres 2015/2016 insgesamt 128 Schüler_innen der 9. Jahrgangsstufe im Bundesland Hamburg zur Bearbeitung vorlegt. Hierbei fand eine systematische Stichprobenziehung statt: Insgesamt wurden 7 Klassen, zweier Gymnasien und zweier Stadtteilschulen befragt. Beide Gymnasien bzw. die beiden Stadtteilschulen unterschieden sich voneinander bzgl. ihres Sozialindexes [26].

Von den 128 befragten Schüler_innen haben n = 116 Schüler_innen die Aufgabe „Weltraumspaziergang“ bearbeitet. Mit Hilfe dieser 116 Schülertexte wurden in einem deduktiv-induktiven Verfahren [16][20] zwei Kategoriensysteme entwickelt, die es ermöglichen, jeden Schülertext in ein Koordinatensystem mit den Dimensionen fachliche und sprachliche Qualität einzuordnen. Jede Dimension wurde aus mehreren Kategorien aufgebaut, die wiederum aus mehreren gestuften Ausprägungen (Subkategorien) bestehen. Entsprechend dieser Stufung ist jeder Ausprägung ein Score (Punktewert) zugewiesen. Beim Codiervorgang wird jedem Schülertext in jeder Kategorie genau eine Ausprägung zugewiesen. Hierdurch erhält jeder Schülertext in jeder Kategorie einen eindeutigen Score. Anschließend werden die Scores jedes Kategoriensystems aufsummiert. Hierdurch erhält man einen Summenscore für die Dimensionen fachlicher Qualität und einen für die sprachliche Qualität der Antwort. Beide Kennwerte erlauben es, die Aufgabenlösung eines Schülers in einem Koordinatensystem zu verorten.

Das Kategoriensystem für die fachliche Qualität eines Schülertextes basiert auf den von Kang et al. mit Verweis auf Braaten und Windschitl entwickelten Kategorien zur Codierung von schriftlichen Schülererklärungen zu physikalischen Sachverhalten [4][12]. Diese Kategorien wurden unter anderem um Aspekte der sog. SOLO-Taxonomie induktiv erweitert [3]. Das Kategoriensystem für die sprachliche Qualität basiert auf dem von der FörMig-Initiative entwickelten Raster zur Beobachtung und Analyse bildungssprachlicher Fähigkeiten von Schüler_innen im natur- und sozialwissenschaftlichen Unterricht für die Sprachhandlung „Erklären“ [18]. Teile des aktuellen Entwurfs dieses Rasters [23] wurde für die Kategoriensystementwicklung übernommen, auf Grundlage der erhobenen Schülertexte konkretisiert und um Aspekte wie z. B. „deiktische Elemente“ [14] und „fachsprachliche Kollokationen“ [30] erweitert.

3.3. Auswahl von Kandidaten für kontrastierende Schülertexte

Im Anschluss an die Entwicklung der Kategoriensysteme erfolgte die Codierung des Gesamtmaterials sowie eine erste Vorauswahl von Prototypen für

[2] Für das zur Verfügung stellen dieser Aufgabe möchten wir uns bei Herrn W. herzlich bedanken.

kontrastierende Schülertexte. Hierzu wurden der Wertebereich des Summenscores für die fachliche Qualität eine Schülertextes (Fachscore) und der Wertebereich des Summenscores für die sprachliche Qualität eines Schülertextes (Sprachscore) in 3 ungefähr gleichgroße Teile zerlegt. Hierdurch wurde das gedachte Koordinatensystem in eine 3x3 Matrix überführt. Die Zellen dieser Matrix wurden nun soweit möglich mit jeweils 3 Schülertexten befüllt (siehe Abb. 4). Dabei galt die zusätzliche Auswahlregel, dass die Schülertexte ungefähr den gleichen Fachscore in jeder Matrixspalte und ungefähr den gleichen Sprachscore in jeder Matrixzeile aufweisen.

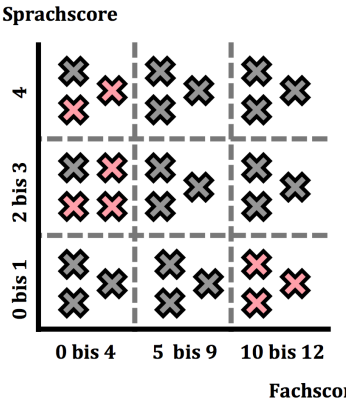

Abb.4: Einordung von Schülertexten in eine 3x3 Matrix. Die roten Kreuze repräsentieren hierbei systematisch abgewandelte Schülertexte.

Bei Zellen, für die es zunächst mehr als 3 passende Optionen gab, wurden diejenigen Texte bevorzugt, die eine größere Textlänge aufwiesen. Für Zellen die zunächst nicht vollständig gefüllt werden konnten, wurden bis dato noch nicht ausgewählte Schülertexte systematisch abgewandelt, um sie für die jeweilige Matrixzelle anzupassen.

Anschließend wurden neben den zuvor 20 + 8 ausgewählten ersten Kandidatentexten zusätzlich 17 weitere Schülertexte aus Erhebung zufällig ausgewählt und durch eine zweite Rater_in codiert. Hierdurch wurden insgesamt 32 % des Originalmaterials doppelt codiert. Als Maß für die Interraterreliablität wurde Krippendorffs-Alpha für ordinale Daten gewählt [15]. Unterschiedliche Codierungen wurden intensiv diskutiert, um ein gemeinsames Verständnis der Kategorien und deren Anwendung zu erreichen und um das Kategoriensystem zu verbessern (vor der Diskussion: $0.12 \leq \alpha \leq 0.73$; nach der Diskussion: $0.88 \leq \alpha \leq 0.92$).

Im Anschluss an die Zweitcodierung wurden aus bis dahin 28 Textoptionen 8 für kontrastierende Schülertexte ausgewählt. Dabei wurden je zwei Texte aus jeder Ecke der in Abb. 4 dargestellten 3x3 Matrix ausgewählt, bei denen Erst- und Zweitrater_in in allen Kategorien identische Ausprägungen zugewiesen hatten.

3.4. Auswahl einer Komposition 4 kontrastierender Schülertexte

Aus den 8 verbleibenden Schülertexten ließen sich insgesamt 16 verschiedene Kompositionen aus 4 kontrastierenden Schülertexten bilden. Im letzten Schritt der Instrumentenentwicklung galt es daher die Frage zu klären, welche dieser möglichen Kompositionen für das Think-Aloud-Instrument die geeignetste ist. Hierzu wurden 6 weitere Rater_innen ausgebildet und folgende Auswahlkriterien festgelegt:

1. Bei welcher der möglichen Kompositionen lassen sich die Kontraste, die in der ersten Codierphase mit zwei Ratern ermittelt worden war, reproduzieren?
2. Bei welcher der möglichen Komposition lassen sich die von Erst- und Zweitrater_in zugewiesenen Ausprägungen der einzelnen Kategorien beider Dimensionen am ehesten reproduzieren?

Um die Aussagekraft dieses Ratings zu erhöhen codierte jede der 6 Rater_innen neben den 8 Kandidatentexten 5 weitere zufällig ausgewählte Schülertexte. Außerdem wurde von einer anschließenden Diskussion unterschiedlicher Codierungen bewusst abgesehen. Für die Auswertung des Ratings wurden folgende parameterfreien Maße bzw. Methoden ausgewählt:

- Kendall's W [13] als Maß für die Konkordanz der Urteile der 6 Rater_innen bei der Codierung der 8 Schülertextoptionen.
- Page's L-Test [21] zur Überprüfung der Vereinbarkeit zwischen den theoretisch erwarteten Trends (zu reproduzierende Kontrastierung) und dem Ergebnis des Ratings.
- Konsenskoeffizient Ξ[3] als Maß für die Raterübereinstimmung bei einem bestimmten Schülertext hinsichtlich einer bestimmten Ausprägung einer Kategorie.

Es zeigte sich eine sehr starke Konkordanz [25, S. 767] der Urteile ($0.72 \leq W \leq 0.94$; $p=0.01$) und in allen 16 Möglichkeiten erwartungsgemäße Kontraste [21, S. 223] sowohl auf der fachlichen ($167.5 \leq L \leq 179.5$; $p=0.01$) als auch auf der sprachlichen Qualitätsdimension ($168.5 \leq L \leq 177.5$; $p=0.01$). Ferner ergab sich aus der Berechnung des Konsenskoeffizienten, dass in 91,7 % der Fälle ein Konsens zwischen den 6 Rater_innen hinsichtlich einer bestimmten Ausprägung einer Kategorie besteht und dass in 77,3 % der Fälle dieser Konsens mit den von Erst- und Zweitrater_in zugewiesenen Ausprägungen übereinstimmt.

[3] Der Konsenskoeffizient Ξ ist eine Modifikation des von Einhaus entwickelten Einigkeitskoeffizienten η [6] für die in diesem Rating vorliegende Datenart [8].

4. Zusammenfassung und Ausblick

Insgesamt konnte mit Hilfe des letzten Entwicklungsschritts genau eine „beste“ Komposition von 4 kontrastierenden Schülertexten hinsichtlich der in Abschnitt 3.3 benannten Auswahlkriterien identifizieren. Die Komposition dieser 4 Schülertexte wurde anschließend in Anlehnung an Arras in die Think-Aloud-Aufgaben für Physiklehrkräfte mit zugehörigem Versuchsleitermanual (Hauptstudie) überführt, wie dies in Abschnitt 2 dargestellt ist [1]. Diese Think-Aloud-Aufgaben wurden an zwei Referendaren und einem Lehramtsstudierenden pilotiert und anschließend verbessert.

Aktuell werden mit diesem Testinstrument Physiklehrkräfte von Gymnasien und Stadtteilschulen im Bundesland Hamburg befragt. Die Datenerhebung der Hauptstudie soll im Laufe des Jahres 2016 abgeschlossen werden. Dabei wird eine Zahl von ca. 15 bis 20 teilnehmenden Lehrkräften angestrebt.

Die in der Erhebungssituation entstandenen verbalen Daten werden mit Hilfe eines Diktiergerätes aufgezeichnet und sollen anschließend mit Hilfe protokoll- und inhaltsanalytischen Methoden ausgewertet werden [5][16][20]. Die miterhobenen demographischen Daten der Lehrkräfte aus einem Fragebogen, sowie deren Korrekturnotizen sollen als zusätzliche Auswertungs- und Interpretationshilfe dienen. Angestrebt wird dabei eine Typisierung, mit deren Hilfe sich die in Abschnitt 2 genannten Forschungsfragen beantworten lassen.

5. Literatur

[1] Arras, U. (2007). *Wie beurteilen wir Leistung in der Fremdsprache?. Strategien und Prozesse bei der Beurteilung schriftlicher Leistungen in der Fremdsprache am Beispiel der Prüfung Test Deutsch als Fremdsprache (Test-DaF)*.Gunter Narr Verlag Tübingen.

[2] Bach, S. (1984). *Systematische und empirische Untersuchung über das Verhältnis von Umgangssprache und Fachsprache im gymnasialen Physikunterricht*. Dissertation. Universität Hamburg.

[3] 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.

[4] Braaten, M., & Windschitl, M. (2011). Working Toward a Stronger Conceptualization of Scientific Explanation for Science Education. *Science Education*, 95 (4), 639-669.

[5] Chi, M.T.H. (1997). Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. *The Journal of the Learning Sciences* , 6 (3), 271-315.

[6] Einhaus, E. (2007). *Schülerkompetenzen im Bereich Wärmelehre. Entwicklung eines Testinstruments zur Überprüfung und Weiterentwicklung eines normativen Modells fachbezogener Kompetenzen*. Logos Verlag Berlin.

[7] 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.

[8] Feser, M.S. (in Vorbereitung). *Fachliche und sprachliche Urteilskriterien von Physiklehrkräften (Arbeitstitel)*. Dissertation, Universität Hamburg.

[9] Freie und Hansestadt Hamburg. Behörde für Schule und Berufsbildung (2011). *Bildungsplan. Gymnasium. Sekundarstufe I. Physik*. Freie und Hansestadt Hamburg.

[10] Freie und Hansestadt Hamburg. Behörde für Schule und Berufsbildung (2014). *Bildungsplan. Stadtteilschule. Jahrgangsstufe 7-11. Physik*. Freie und Hansestadt Hamburg.

[11] 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.

[12] 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.

[13] Kendall, M.G. (1948). *Rank Correlation Methods*. Griffin.

[14] Kniffka, G., & Siebert-Ott, G. (2012). *Deutsch als Zweitsprache. Lehren und Lernen*. 3 ed. Verlag Ferdinand Schöningh.

[15] Krippendorff, K. (2004). *Content Analysis. An Introduction to its methodology*. 2 ed. SAGE Publications Inc..

[16] Kuckartz, U. (2016). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung*. Beltz Verlag.

[17] Kulgemeyer, C. (2010). *Physikalische Kommunikationskompetenz. Modellierung und Diagnostik*. Logos Verlag Berlin.

[18] 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.

[19] 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.
[20] Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. Beltz Verlag.
[21] Page, E.B. (1963). Ordered Hypotheses for Multiple Treatments: A Significance Test for Linear Ranks. *Journal of the American Statistical Association*, 58 (301), 216-230.
[22] Rincke, K. (2007). *Sprachentwicklung und Fachlernen im Mechanikunterricht. Sprache und Kommunikation bei der Einführung in den Kraftbegriff*. Logos Verlag Berlin.
[23] Roth, H.J., Lengyel, D., & Reich, H.H. (2011). *Sprachhandlungen Erklären, Berichten und Argumentieren. Manual zu den Auswertungsrastern. Entwurfsfassung*. FÖRMIG AG SEK I.
[24] Schleppegrell, M.J. (2004). *The Language of Schooling. A Functional Linguistics Perspective*. Routledge.
[25] Schmidt, R.C. (1997). Managing Delphi Surveys Using Nonparametric Statistical Techniques. *Decision Sciences*, 28 (3), 763-774.
[26] Schulte, K., Hartig, J., & Pietsch, M. (2014). Der Sozialindex für Hamburger Schulen. In D. Fickermann, & N. Maritzen (Eds.), *Grundlagen für eine daten- und theoriegestütze Schulentwicklung. Konzeption und Anspruch des Hamburger Instituts für Bildungsmonitoring und Qualitätsentwicklung (ifBQ)* (pp. 67-80). Waxmann.
[27] Strate, W. (2014). *Physik Schulaufgaben, Übungen. 7.-10. Klasse Gymnasium*. Amazon Distribution GmbH.
[28] Südkamp, A., Möller, J., & Kaiser, J. (2012). Accuracy of Teachers' Judgements of Students' Academic Achievement: A Meta-Analysis. *Journal of Educational Psychology*, 104 (3), 743-762.
[29] 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.
[30] Tajmel, T. (2011). Wortschatzarbeit im mathematisch-naturwissenschaftlichen Unterricht. *i-de*, 35 (1), 83-93.
[31] Thonhauser, J. (2008). Warum (neues) Interesse am Thema ‚Aufgaben'? . In J. Thonhauser (Ed.), *Aufgaben als Katalysatoren von Lernprozessen. Eine zentrale Komponente organisierten Lehrens und Lernens aus der Sicht von Lernforschung, Allgemeiner Didaktik und Fachdidaktik* (pp. 13-26), Münster: Waxmann.
[32] van Someren, M.W., Barnard, Y.F., & Sandberg, J.A.C. (1994). *The Think Aloud Method. A practical guide to modelling cognitive processes*. Academic Press.