Science Education
Review Letters

Do Linguistic Features Influence Item Difficulty in Physics Assessments?

Dietmar Höttecke[a*], Markus Sebastian Feser[a], Lena Heine[b], Timo Ehmke[c]
[a] Universität Hamburg, Faculty of Education, Physics Education, Germany
[b] Ruhr-Universität Bochum, Faculty of Philology, German Philology, Language Education & Multilingualism, Germany
[c] Leuphana University of Lüneburg, Faculty of Education, Institute of Educational Sciences, Germany

[*] Corresponding author: dietmar.hoettecke@uni-hamburg.de

Received 7th March 2018, Accepted 14th May 2018

Abstract
This paper addresses the question if and to which extent linguistic surface features of test items in a physics assessment affect item difficulty. In an experimental study, linguistic features of test items in physics were varied systematically on three levels based on a heuristic model of linguistic demands. The results show that item difficulty can be predicted by linguistic features, but only for a limited number of items and not in a consistent way.

Keywords
linguistic demands, item difficulty, physics

Theoretical background and state of research

In order to be academically successful, students have to develop linguistic competences beyond the use of everyday language registers (Schleppegrell, 2004). Teachers have to support their students' competence development, for instance by providing texts (e.g. textbooks, worksheets) at an appropriate linguistic level, which allows for understanding a specific content as well as for the acquisition of new linguistic repertoires. Furthermore, texts in assessments have to be comprehensible for all students in order to avoid construct irrelevant variation and item bias.

Text comprehension and readability research (a good overview can be found in Beinborn, 2016) has shown that a text's level of difficulty is not only affected by its lexical and syntactic structures, but strongly depends on the level of cognitive structuring, level of cohesion, semantic redundancy, as well as the previous knowledge of the reader. The relevance of these factors has hitherto been demonstrated by text comprehension studies in science (e.g., Diebold & Waldron, 1988; Deppner, 1989; Sumfleth & Schüttler, 1995; Starauschek, 2006; Cromley et al., 2010). It is obvious to assume that simplification of texts increases its comprehensibility, but findings hitherto are not consistent. It has been demonstrated that high textual cohesion, which was expected to increase a text's comprehensibility, can create comprehension difficulties for students with a high level of prior

knowledge (Härtig & Kohnen, 2017; Kohnen et al., 2017).

Research on science test item difficulty has addressed linguistic characteristics of test items alongside other features that are known to influence item difficulty, such as positional effects, visual information, openness of response formats, length of answering options in closed items, degree of abstractness, domain-specific cognitive processes, subject-specific previous knowledge, subject-specific terminology, complexity of information etc. The *Progress in International Reading Literacy Study* (PIRLS) (Stubbe, 2011) reveals that rather slight linguistic variation among translations into German caused measurable differences in difficulty. If linguistic features affect item difficulty, an item bias induced by language might be a general problem in (large-scale) assessments (El Masri et al., 2016).

Test item research with a focus on linguistic features in physics in particular and in science or mathematics in general is rare: In science, it has been shown that the use of technical terms causes difficulty (Stiller et al., 2016). The same holds for test items in biology (Schmiemann, 2011) and mathematics (Fischer-Hoch et al., 1997). Cassels and Johnston (1984) simplify one of two identical matched items (lexis, use of negation and text length) and show an effect on the rate of solutions. Bird and Welford (1995) follow a similar approach. They vary lexis, text length, syntax and tense, and also show effects on item difficulty, and that the effects are greater for L2 learners (Botswana) than for L1 learners. In a similar study, Prophet and Badede (2009) show positive effects of linguistic simplification. They also demonstrate that an extended reduction of text length leads to an opposite effect. Llosa et al. (2016) finally fail to show any effect on test difficulty for L2 learners, if words which are assumed to be hard to understand were systematically avoided across an assessment. In a meta-study on item bias from Anglophone research (Kieffer et al., 2009), only the use of dictionaries and glossaries has been proven to be effective in order to reduce item bias, while any simplification of the test language (English) was not. For items in mathematics, Haag et al. (2015) failed to show any main effect of linguistic simplification but could demonstrate a slight advantage for students with intermediate language proficiency. This effect was slightly stronger for students using a minority language at home.

As this overview demonstrates, a number of linguistic, textual and content dimensions related to difficulty seem to interact with each other in a complex manner which is presently not fully understood. It thus seems to be necessary to investigate the impact of linguistic features on item difficulty in more detail. In the explorative study presented here we will focus on linguistic difficulty of text-laden physics test items. It will be investigated in how far linguistic surface factors increase the difficulty of physics test items for 8th and 9th grade students.

Research question and study design

Two general research questions are in the center of our study: (1) To what degree do high levels of linguistic demands increase the degree of difficulty of the items? We assume that the use of linguistic features producing higher cognitive load on linguistic processing results in more effort to construct meaning. The probability of solving a particular item will thus decrease with an increasing level of linguistic demands of the item stem. However, we do not know how big this effect actually is. (2) Does students' language proficiency predict their test performance in a physics assessment and if so to which extent?

In order to test these assumptions on an exploratory level, two studies have been conducted. In a first study, Physics tasks were administered as part of a broader research design incorporating a number of different school subjects (Physics, Maths, Music, PE, German, Leiss et al., 2017) in a multi-matrix design (N = 601, grade 7/8 from 29 German upper secondary classes, approx. N = 200/item). We operationalize the linguistic variation according to three principles that can be assumed to create linguistic difficulty in language comprehension. The principles were synthesized from evidence from psycholinguistic, language testing and readability research (Heine et al., to appear) and focus on lexical, syntactic and semantics dimensions:

1. Structural complexity of linguistic elements: higher structural complexity increases the cognitive demand of linguistic processing (e.g., simple and short sentences are easier to process than long sentences with embedded clauses).

2. Semantic transparency of form-meaning units: the less straightforward meaning can be mapped onto a linguistic element the more dif-

ficult it is (e.g., idiomatic expressions are more difficult than non-idiomatic expressions).

3. Frequency of linguistic elements in language use: the less frequent an element is, the less can it be assumed to be part of language users' knowledge base and the less automated it is. As a consequence, the strain of cognitive processing is increased (e.g. "edifices" is less frequent than "house").

These dimensions can be addressed by variation of certain linguistic surface phenomena and relate to the fact that academic language is harder to understand for students than everyday language. Multiple-choice-answers were not varied linguistically.

The expected effect that linguistically less complex items were solved with a higher probability was detected only for a minority of items on a significant level. This result led to a second study which is reported here and in which a more rigidly controlled operationalization of the linguistic and task dimensions was carried out. For this study, 6 physics items were developed following a series of steps:

1. Two physics education researchers developed six preliminary multiple-choice items. These items addressed typical content of German curricula for introductory physics classes (simple electric circuits, magnetic phenomena). The content was recommended by two experts from pre- and in-service physics teacher training, because of its high probability to be actually taught in class 7/8. Each item stem was followed by one multiple-choice item.

2. The item stems were systematically modified on three levels of linguistic demands according to the model briefly described above. This was done by a group of five experts from linguistics (for details see Heine et al., to appear). Multiple-choice options were not varied linguistically, but held constant across levels of linguistic demands of item stems.

3. These 6x3 items were discussed and rechecked by the two physics education researchers to ensure that the physics content did not vary across levels of linguistic demands.

4. All items were piloted (observation of partner work + interview) in order to ensure their general coherence and comprehensibility for students (N = 18).

5. Finally, items were carefully revised according to the results of the pilot study.

Shortened item stem	Translation into English
Level 1: Frau Schröder ist Physiklehrerin. Sie zeigt ihrer Klasse ein Experiment. Sie baut einen elektrischen Stromkreis. Sie nimmt dazu eine Batterie, eine Glühlampe und mehrere Kabel. Sie lässt auch Platz für genau einen Stab im Stromkreis. Frau Schröder hat verschiedene Stäbe. Die Stäbe sind aus Holz, Graphit, Glas, Kupfer und Gummi. Sie haben die gleiche Form. Frau Schröder setzt die Stäbe nacheinander in den Stromkreis. Die Klasse sieht: Die Glühlampe leuchtet mit manchen Stäben hell. Sie leuchtet mit anderen Stäben aber nur schwach. Mit manchen Stäben leuchtet sie gar nicht. Frau Schröder fragt die Klasse: „Welche Frage kann man mit dem Experiment beantworten?" Fünf Schülerinnen und Schüler antworten. […]	**Level 1:** Mrs. Schröder is a physics teacher. She shows an experiment to her class. She builds an electrical circuit. For this she uses a battery, a light bulb and several wires. She also leaves space for exactly one rod in the circuit. Mrs. Schröder has different rods. The rods are made of wood, graphite, glass, copper and rubber. They have the same shape. Mrs. Schröder puts one rod after the other into the circuit. The class sees: The light bulb glows with some rods. It glows with other rods but only weakly. With some rods it does not glow at all. Mrs. Schröder asks the class: "What question can be answered with the experiment?" Five pupils answer. […]
Level 3: Im Rahmen eines Experiments, das sie ihrer Klasse vorführt, baut die Physiklehrerin Frau Schröder einen elektrischen Stromkreis auf, wozu eine Batterie, eine Glühlampe sowie mehrere Kabel zum Einsatz kommen. An lediglich einer Stelle kann ein zusätzlicher Stab eingesetzt werden. Verschiedene gleich geformte Stäbe aus Holz, Graphit, Glas, Kupfer und Gummi stehen Frau Schröder zur Verfügung. Als sie diese einen nach dem anderen in den Stromkreis einsetzt, ist ersichtlich, dass, je nachdem, welche Stäbe verwendet werden, die Glühlampe mal hell, schwach oder auch mal gar nicht leuchtet. Nachdem die Lehrerin die Aufgabe an die Klasse gerichtet hat zu beantworten, welche Frage man mit dem Experiment beantworten könne, bieten fünf Schülerinnen und Schüler eine Antwort an: […]	**Level 3:** In the course of an experiment she is presenting in her class, the physics teacher Mrs. Schröder builds an electrical circuit, using a battery, a light bulb, and several wires. An additional rod can be used in only one place. Various equally-shaped rods made of wood, graphite, glass, copper and rubber are available to Mrs. Schröder. When one of these rods is inserted into the circuit, it becomes visible that, depending on which rods are used, the light bulb is bright, weak or sometimes not glowing at all. After the teacher asked the class to answer what question one could answer with the experiment, five students offer an answer: […]

Fig. 1: Shortened item stem of item P3 on level 1 and 3 of linguistic demands. The task is to decide which of the following five student responses is correct ("What materials are electric conductors or non-conductors?"*).*

Item	Relative frequency of correct responses			Oneway ANOVA
	low language demands	medium language demands	high language demands	
P1	0,17	0,36	0,36	F(2, 436.22) = 15.53, p < .01 (Welch)
P2	0,44	0,42	0,43	F(2, 673) = 0.05, n.s.
P3	0,70	0,55	0,57	F(2, 443.85) = 6.96, p < .01 (Welch)
P4	0,08	0,09	0,09	F(2, 673) = 0.92, n.s.
P5	0,36	0,36	0,35	F(2, 673) = 0.06, n.s.
P6	0,28	0,23	0,23	F(2, 443.74) = 0.82, n.s.

Tab. 1: Relative frequency of correct responses and explanation of variance for each of the physics items

The final instrument obtained in this way consisted of 6 physics items on 3 levels of linguistic demands. The test language was German. Fig. 1 illustrates the span of linguistic variation (level 1 and 3, item P3); for reasons of illustration, an approximate English translation is presented alongside the German original which attempts to mimic the German structures and is therefore not fully idiomatic.

Additionally, a reduced c-test was administered (2 x 30 items) which deviated from the canonical c-test concept (Grotjahn, 2006), in order to measure the students' general language competence in German. Further questions about the students' linguistic and migration background, gender, age, their self-estimation of language proficiency in German (talking, reading, understanding, writing), grades in several subjects as well as cultural capital (number of books at home) were asked.

The whole sample covers N = 1346 German secondary-school students from 17 schools (50.6% female, age mean = 14.0, 33.6% migration background). Test items were rotated in a matrix-design across six test booklets, which let to 220-227 answers per item. For results for all participating subjects see Schwippert et al. (submitted).

Results

A one-way ANOVA was calculated with linguistic demands as factor and test scores of physics items as response variable. The calculation was done for each of the six physics items in order to investigate research question (1) (Tab. 1). The relative frequency of correct responses for item P1-P6 are presented for each of the levels of linguistic demands from low to high. It turns out that the coherence between test scores and linguistic demands is general low (research question 1). Variance of each of the items could be explained in a significant manner only for item P1 and P3, while only P3 is indicating the effect as expected. Item difficulty of P4 is too high and thus does not need to be considered further because of its bottom effect and lack of generate variance. P2, P5 and P6 show the expected effect slightly, but not on a significant level. These results indicate that a reduction of the degree of academic language features does not reduce item difficulty in a clear and coherent manner. Such a reduction may even lead to an increasing difficulty as shown by item P1.

In order to answer research question (2) bivariate correlations and multiple regression models were calculated (Tab. 2). Results presented here are based on a series of regression models, which led to the identification of significant predictors (method: forward) for each of the physics items. This analysis shows that the level of linguistic demands significantly predicts test scores for P1 and P3 only. Students' language proficiency in German based on c-test data predicts test scores for P1, P3, P5 and P6 as expected, but not for P2. Thus, this study presents evidence for the assumption that language proficiency predicts test performance (research question 2).

Discussion

There is an ongoing discussion to which extent linguistic characteristics of test items cause item difficulty and whether students benefit from a linguistic simplification of texts for teaching and learning in science in general and in physics in particular. Furthermore, it is not clear if linguistic characteristics of test items cause construct-irrelevant variance. Physics item stems were varied systematically on three levels of linguistic demands. We did not find a coherent main effect of linguistic demands of academic language on item difficulty. One of the items (P1)

		r (biv., Pearson)	unstandardized B	unstandardized SE	standardized BETA	p	R^2 (corr.)
P1	c-test	.197**	.006	.001	.182	<.001	
	level of linguistic demands	.172**	.097	.023	.175	<.001	
	grade in Math	–.153**	–.23	.010	–.107	<.1	.080
P2	migration background	–.259**	–.127	.027	–.200	<.001	
	number of books at home	.224**	.050	.013	.162	<.001	
	reading German (self-estimated)	–.050	–.084	.032	–.109	<.01	.108
P3	c-test	.397**	.013	.001	.380	<.001	
	level of linguistic demands	–.113**	–.064	.024	–.107	<.01	.153
P4	c-test	–.157**	–.004	.001	–.216	<.001	
	grade in German	–.051	–.035	.010	–.261	<.001	
	grade in Physics	.001	.019	.009	.153	<.1	.041
P5	c-test	.188**	.007	.001	.203	<.001	.039
P6	c-test	.109**	.004	.001	.140	<.001	.018

*Table 2: Results of multiple regression analysis and bivariate correlations (**: $p \leq 0.01$, *: $p \leq 0.05$)*

even shows a reversal effect and indicates that it is possible that a low level of linguistic demands might even cause difficulty. Even though the variation of linguistic demands across the three levels should not affect the content, it might nevertheless have happened that the content was presented in a more demanding way on level 1. If for instance a series of main clauses (level 1) requires a higher level of inferences for reconstructing a mental model compared to a sequence of main and subordinate clause on higher levels, a higher cognitive demand on a low level of linguistic demands compared to higher levels might be a consequence. This interpretation is in line with the general assumption that the construction of a mental model of a text is a complex interaction of traits of the text with traits of a person. As a consequence, the potential of a systematic variation of linguistic surface phenomena towards affecting item difficulty might be limited. Further research in this field is needed.

Our study is characterized by some limitations: Several factors that influence item difficulty have not been controlled e.g. content areas, cognitive process or response formats. This will be taken into account in future research. Nevertheless, the preliminary evidence presented here is in line with findings from Kieffer et al. (2009), Haag et al. (2015) and others that the impact of linguistic surface characteristics of academic language on text difficulty in general and item difficulty in particular, if it really does exist, is rather small. Nevertheless, future studies have to consider further sources of variance like cognitive activity or complexity of information which has to be processed more systematically.

References

Beinborn, L. (2016). Predicting and manipulating the difficulty of text-completion exercises for language learning. Online dissertation, TU Darmstadt, http://tuprints.ulb.tu-darmstadt.de/5647/1/DissertationBeinborn_publishedVersion_September20_online.pdf, last access March 02, 2018

Bird, E. & Welford, G. (1995). The effect of language on the performance of second-language students in science examinations. *IJSE, 17* (3), 389–397.

Cassels, J. R. T. & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education, 61* (7), 613–615.

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Reading comprehension of scientific text: A domain-specific test of the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 102*, 687–700.

Deppner, J. (1989). *Fachsprache der Chemie in der Schule. Empirische Untersuchung zum Textverständnis und Ansätze zur sprachlichen Förderung türkischer und deutscher Schülerinnen und Schüler.* Heidelberg: Julius Groos Verlag.

Diebold, T.J. & Waldron, M.B. (1988). Designing Instructional Formats: The Effects of Verbal and Pictorial Components on Hearing-Impaired Students' Comprehension of Science Concepts. *American Annals of the Deaf, 133* (1), 30–35.

El Masri, A. H., Baird, J.-A. & Graesser, A. (2016). Language effects in international testing: the case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice, 23* (4), 427–455.

Fisher-Hoch, H., Hughes, S., & Bramley, T. (1997). *What makes GCSE examination questions difficult? Outcomes of manipulating difficulty of GCSE questions.* Paper presented at the British Educational Research Association Annual Conference, University of York, http://www.leeds.ac.uk/educol/docu-

ments/000000338.htm (15.02.2018).

Grotjahn, R. (Hrsg.) (2006). *Der C-Test: Theorie, Empirie, Anwendung.* Frankfurt a.M.: Peter Lang Verlag.

Haag, N., Heppt, B., Roppelt, A., & Stanat, P. (2015). Linguistic simplification of mathematics items: effects for language minority students in Germany. *Eur J Psychol Educ, 30*, 145–167.

Härtig, H. & Kohnen, N. (2017). Die Rolle der Termini beim Lernen mit Physikschulbüchern. In B. Ahrenholz, B. Hövelbrinks & C. Schmellentin (Hrsg.), *Fachunterricht und Sprache in schulischen Lehr-/Lernprozessen* (S. 55–72). Tübingen: Narr.

Heine, L., Domenech, M., Otto, L., Neumann, A., Krelle, M.,Leiß, Dominik, & Höttecke, D. (to appear). Modellierung sprachlicher Anforderungen in Testaufgaben verschiedener Unterrichtsfächer: Theoretische und empirische Grundlagen. *Zeitschrift für angewandte Linguistik*.

Kieffer, M. J., Lesaux, N. K., Rivera, M. & Francis, D. J. (2009). Accommodations for English Language Learners Taking Large-Scale Assessments: A Meta-Analysis on Effectiveness and Validity. *Review of Educational Research, 79* (3), 1168–1201.

Kohnen, N., Härtig, H., Bernholt, S. & Retelsdorf, J. (2017). Leichte Sprache im Physikunterricht. In B. M. Bock, U. Fix & D. Lange (Hrsg.), *„Leichte Sprache" im Spiegel theoretischer und angewandter Forschung* (S. 337–341).

Leiss, D., Domenech, M., Ehmke, T. & Schwippert, K. (2017). Schwer – schwierig – diffizil: Zum Einfluss sprachlicher Komplexität von Aufgaben auf fachliche Leistungen in der Sekundarstufe I. In D. Leiss, M. Hagena, A. Neumann, & K. Schwippert (Hrsg.), Mathematik und Sprache. *Empirischer Forschungsstand und unterrichtliche Herausforderungen* (S. 99-125). Münster [u.a.]: Waxmann.

Llosa, L., Lee, O., Jiang, F., Haas, A., O'Connor, C., Van Booven, C. D. & Kieffer, M. J. (2016). Impact of a Large-Scale Science Intervention Focused on English Language Learners. *American Educational Research Journal, 53* (2), 395–424.

Prophet, R. B. & Badede, N. B. (2009). Language and student performance in Junior Secondary Science examinations: The case of second language learners in Botswana. *International Journal of Science and Mathematics Education, 7*, 235–251.

Schleppegrell, M. J. (2004). *The language of schooling. A functional linguistics perspective.* London: Lawrence Erlbaum Associates.

Schmiemann, P. (2011). Fachsprache in biologischen Testaufgaben. *Zeitschrift für Didaktik der Naturwissenschaften, 17*, 115-136.

Schwippert, K., Leiss, D., Ehmke, T., Höttecke, D., Heine, L., & Neumann, A. (submitted). Die Arbeitsgruppe Fach und Sprache (AG FuS): Empirische Fundierung der Untersuchung von sprachsensiblen Fachaufgaben.

Starauschek, E. (2006). Der Einfluss von Textkohäsion und gegenständlichen externen piktoralen Repräsentationen auf die Verständlichkeit von Texten zum Physiklernen. *Zeitschrift für Didaktik der Naturwissenschaften, 12*, 127-157.

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education, 41* (5), 721-732.

Stubbe, T.C. (2011). How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments. *Educational Research and Evaluation, 17* (6), 465-481.

Sumfleth, E. & Schüttler, S. (1995). Linguistische Textverständlichkeitskriterien. Helfen Sie bei der Darstellung chemischer Inhalte? *Zeitschrift für Didaktik der Naturwissenschaften, 1*, 55–72.