International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: http://www.tandfonline.com/loi/tsed20

Physics knowledge of first semester physics students in Germany: a comparison of 1978 and 2013 cohorts

David Buschhüter, Christian Spoden & Andreas Borowski

To cite this article: David Buschhüter, Christian Spoden & Andreas Borowski (2017): Physics knowledge of first semester physics students in Germany: a comparison of 1978 and 2013 cohorts, International Journal of Science Education, DOI: 10.1080/09500693.2017.1318457

To link to this article: http://dx.doi.org/10.1080/09500693.2017.1318457

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION, 2017
https://doi.org/10.1080/09500693.2017.1318457

Physics knowledge of first semester physics students in Germany: a comparison of 1978 and 2013 cohorts

David Buschhüter[a], Christian Spoden[b] and Andreas Borowski[a]

[a]Physics Education, Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany; [b]Department of Research Methods in Education, Friedrich-Schiller-University Jena, Jena, Germany

ABSTRACT

Over the last decades, the percentage of the age group choosing to pursue university studies has increased significantly across the world. At the same time, there are university teachers who believe that the standards have fallen. There is little research on whether students nowadays demonstrate knowledge or abilities similar to that of the preceding cohorts. However, in times of educational expansion, empirical evidence on student test performance is extremely helpful in evaluating how well educational systems cope with the increasing numbers of students. In this study, we compared a sample of 2322 physics freshmen from 2013 with another sample of 2718 physics freshmen from 1978 at universities in Germany with regard to their physics knowledge based on their results in the same entrance test. Previous results on mathematics knowledge and abilities in the same sample of students indicated that there was no severe decline in their average achievement. This paper compares the physics knowledge of the same two samples of students. Contrary to their mathematics results, their physics results showed a substantial decrease in physics knowledge as measured by the test.

ARTICLE HISTORY
Received 28 June 2016
Accepted 9 April 2017

KEYWORDS
University physics; entrance test; generational comparison

Educational expansion

In the 1960s, Kingsley Amis commented on the consequences of the massification of higher education:

> I wish I could have a little tape-and-loudspeaker arrangement sewn into the binding of this magazine to be triggered off by the light reflected from the reader's eyes on to this part of the page, and set to bawl out at several bells: MORE will mean WORSE. (Amis, 1960, p. 8)

In a miscellany edited by Tapper and Palfreyman (2005b), authors from a range of countries presented comparative perspectives on their understanding of mass higher education. They concluded that the expansion had been welcomed in many countries 'in spite the occasional reservation – perhaps most infamously expressed in Kingsley Amis' 'more means worse' (Tapper & Palfreyman, 2005a, p. 247). From an Australian perspective, Duke (2005) reported that the suggestion 'more has meant worse' has 'muted' (p. 25)

CONTACT David Buschhüter david.buschhueter@uni-potsdam.de Physics Education, Institute of Physics and Astronomy, University of Potsdam, Karl-Liebknecht-Straße 24/25, 14476 Potsdam, Germany

on the side of universities and politics, both being motivated to defend themselves against media attacks concerning falling standards.

However, research on self-theories of university staff has shown evidence supporting the idea that university teachers might, to a large proportion, still be in favour of Amis' idea. Yorke and Knight (2004) asked 72 university staff members in the UK about their agreement with statements such as 'Today, students in higher education are just as talented as they used to be'. This statement is supposed to be an implicit inversion of Kingsley Amis' 'MORE will mean WORSE', and it is interesting to notice that the majority (59%) of the university staff members tended towards rejecting the statement (Yorke & Knight, 2004).

There is certainly a need to investigate empirically the extent to which Amis' statement is actually true. As it is not possible to vary entry rates experimentally, it is reasonable to study an educational system that underwent educational expansion and investigate how the average educational achievement has changed over the years. Therefore, we investigated the extent to which the teachers' belief 'Today, students in higher education are not as talented as they used to be' might be true. It needs to be mentioned that talent is not to be confused with achievement. University teachers usually never measure talent directly, but witness and interpret student achievement. When university teachers state that students nowadays have less talent, it is probable that this statement is based on their experience in certain domain-specific learning situations. In other words, they think that students nowadays do not match the content-related requirements as well as students did before.

In this paper, we focus on content-related requirements for physics students in Germany. We investigated the extent to which these two cohorts' content-related knowledge or abilities had changed over the last decades. This type of research is rare, as historic data from older cohorts is necessary for such comparison. Fortunately, in 1978 a national entrance test was conducted in Western Germany with first year undergraduate physics students ($N = 2718$). This test and the corresponding item difficulties (relative frequencies of correct item responses) were published in 1981, allowing us to replicate the study with a cohort in 2013. The booklet consisted of a physics test and a mathematics test. In a previous analysis, we focused on differences in mathematics knowledge and abilities (Buschhüter, Spoden, & Borowski, 2016). In order to provide a more complete profile, the present paper focuses on the question of how physics freshmen in 1978 and 2013 differed in their physics knowledge with significant relevance to current physics programmes. This example from Germany helps us to understand the extent to which educational systems can meet the challenges of increasing enrolment – a challenge that countries all over the world are currently facing.

In the following sections, we provide evidence for the increase in first semester student enrolment rates in physics between 1978 and 2013 to show that the physics programmes underwent educational expansion. We then provide a brief summary of the mathematics results, showing that there was no empirical evidence to believe that students in 2013 would perform worse compared to those in 1978 regarding mathematics knowledge and abilities as the key requirements for the physics programmes. Subsequently, we show that physics knowledge can also be regarded as part of the content-related requirements and should therefore be compared. As no hypothesis about the results of the comparison could be generated a priori, the subsequent analysis was exploratory.

Previous research

Changes in German freshmen rates

In Western Germany, the proportion of freshmen at universities relative to the population of their age group was approximately 12% in 1980. We assume a similar rate in 1978. In 2013 this rate was about 34% in Germany. The proportion of physics freshmen relative to all university freshmen has been more or less constant with about 3% in 1978 (Western Germany) and 2.7% (Germany) in 2013. As a result, the estimated enrolment rate of physics freshmen students in 2013 was more than twice as high as the estimated rate in 1978 based on data of the Federal Ministry of Statistics (Buschhüter et al., 2016). These rates included students enrolling in physics courses exclusively in order to benefit from the special 'student status' in Germany (e.g. in order to get a discount on public transport). However, these freshmen enrolment rates indicated that the physics programme in Germany is a valid example of a system undergoing educational expansion. Using the results of the national entrance test of mathematics of 1978, we previously provided evidence that this expansion did not generally lead to lower knowledge and abilities in mathematics among physics students (Buschhüter et al., 2016). In the following section, we briefly summarise the results.

Changes in mathematics knowledge and abilities of German physics freshmen of 1978 and 2013

In 1978, the national entrance test for university freshman in physics was used to study the knowledge and abilities of students entering the university (Krause & Reiners-Logothetidou, 1981). Both the mathematics and the physics parts of the test consisted of 47 items. Krause and Reiners-Logothetidou published the test, the manual about how to conduct the test, the scoring rubric and item difficulties (percentages of correct item responses), offering the opportunity for us to replicate the study of 1978 with a sample of physics students in 2013.

In 1978, 2718 physics freshmen took the test, whereas 2322 physics freshmen did in 2013. The results of the mathematics test indicated a minor decline in the mean raw score in 2013 compared to that in 1978. This difference was significant, $t(5038) = 3.20$, $p = .001$, albeit negligibly small in effect size, $d = 0.09$. The results also provided evidence that the students in 1978 outperformed those in 2013 on some items (e.g. items concerning equations), but vice versa on other items (e.g. most items concerning integrals). Figure 1 shows the differences in mean item difficulties between the 1978 and 2013 cohorts depending on the content area of relevance. The results of this study were surprising, as the test should be oriented closer to the curriculum of 1978 than to that in 2013. These results are discussed later in the paper in combination with the physics test results.

It needs to be pointed out that the results for different content areas are based on different numbers of items; thus, the difference in the mathematics overall score between the 1978 and 2013 cohorts is not the average difference in score across these content areas. Also, it should be noted that the students' abilities in different mathematical content areas might have a different impact on their achievement in physics. For example, due to the hierarchical structure of mathematics, a decrease in the ability to solve equations

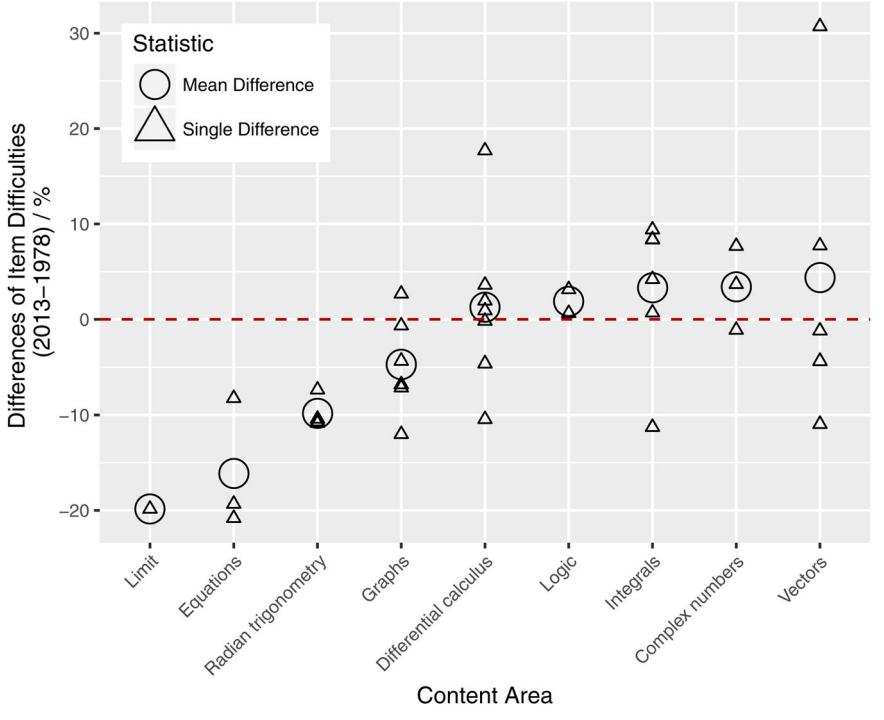

Figure 1. Differences in item difficulties by content area for the mathematics test (Buschhüter et al., 2016, p. 12, with permission of Springer; partial results from Krause & Reiners-Logothetidou, 1981, pp. 305–310).

might result in more serious learning difficulties than a decrease in the ability to use integrals or complex numbers. In that sense, differences in average scores can be misleading when changes on the item level are not investigated. The results on the item level contradict the belief that students nowadays show a lower level of mathematics knowledge and abilities in general.

Mathematics knowledge and abilities represent the most prominent group of content-related requirements. However, the knowledge of physics can also be relevant in the beginning of university studies. In the next section, we present the extent to which physics and mathematics differ in their role as domains of content-related requirements.

Physics knowledge as a domain of requirements

The average decrease in the knowledge or abilities of physics freshmen is relevant when these skills are directly related to the requirements of the university programmes. Albrecht (2011) provided evidence that content-related requirements are among the most important dropout motives in physics programmes in Germany. A large proportion of the students dropping out mentioned mathematical methods or theoretical physics, which demands high mathematics knowledge and abilities, as a reason for their decision to leave the programmes (Albrecht, 2011). Many studies have shown that performance in physics university programmes are correlated with mathematics pre-schooling, knowledge

and abilities (e.g. Hazari, Tai, & Sadler, 2007; Hudson & McIntire, 1977; Long, McLaughlin, & Bloom, 1986; Sadler & Tai, 2001). This does not indicate that there is a direct causal relation; but as physics makes use of mathematical representations and is closely related to mathematics (e.g. Kragh, 2015; Tuminaro, 2004), it is reasonable to assume that lower physics achievement, and eventually, the decision to leave the physics programmes might often be an outcome of a lack of mathematics-related knowledge or abilities (Albrecht, 2011).

In a study with coordinators of college chemistry by Shumba and Glass (1994), the authors found mathematics to be of more importance than the science content. One faculty member stated: 'While basic information in the sciences is important, math through spherical trig and elementary calculus and application of those skills through word problems is more important' (Shumba & Glass, 1994, p. 389).

As shown in commonly used physics textbooks (e.g. Young, Freedman, Ford, & Sears, 2012), it seems likely that physics courses start with the basics of the physics concepts but do not provide sufficient learning opportunities for students to master all the necessary mathematical methods. Obviously, this by no means implies that students' prior knowledge in physics is not an important predictor of physics knowledge assessed at university level. Despite the doubts of some college professors, research found prior-performance or pre-schooling in physics to be predictive for success in physics courses (Halloun & Hestenes, 1985; Sadler & Tai, 2001).

From a constructivist point of view (e.g. Driscoll, 2005; Uzuntiryaki, Boz, Kirbulut, & Bektas, 2010), prior knowledge in physics influences the process of building new knowledge in university physics. Additionally, physics knowledge might also compensate for an individual's lack of mathematics knowledge.

Thus, mathematics and physics knowledge are different types of requirements relevant for physics studies. Prior mathematics knowledge or abilities are important from the curricular point of view, whereas physics prior knowledge is infrequently demanded by the curriculum itself but becomes a significant issue in the construction of new knowledge.

The average decrease in knowledge or abilities of physics freshmen is especially important when the specific knowledge or abilities are related to the requirements of the university programmes. Content matter experts are needed to evaluate whether a particular ability or some kind of knowledge is valid as being necessary or helpful for studying the physics programmes. The test score measuring the extent to which a person fulfils these requirements should also be predictive for learning achievement at the university level. The relevance of the items used in this study was validated in this regard. The next section describes the kind of physics knowledge that was measured by the test in 1978.

Physics knowledge of the national entrance test in 1978

Krause and Reiners-Logothetidou (1981) stated that the content of the entrance test of 1978 was relevant regarding the requirements for the physics university programmes. They also argued that the majority of physics items relate to a term- or definition-related qualitative knowledge that would be important for solving concrete problems or understanding the method of physics (Krause & Reiners-Logothetidou, 1981). We describe the test items in Figure 2.

Figure 2 shows that only a very small proportion of items demanded students to calculate or to plot a graph. The majority of the items required recalling information. Also, the activities 'explain' and 'estimate' are influenced by declarative knowledge as they cover recalling physics knowledge or values from physical quantities. Even the items in the category 'draw' could be interpreted as graphical reproduction. Measuring current in a circuit and drawing a ray diagram for a lens are standard exercises. This is not true for the items

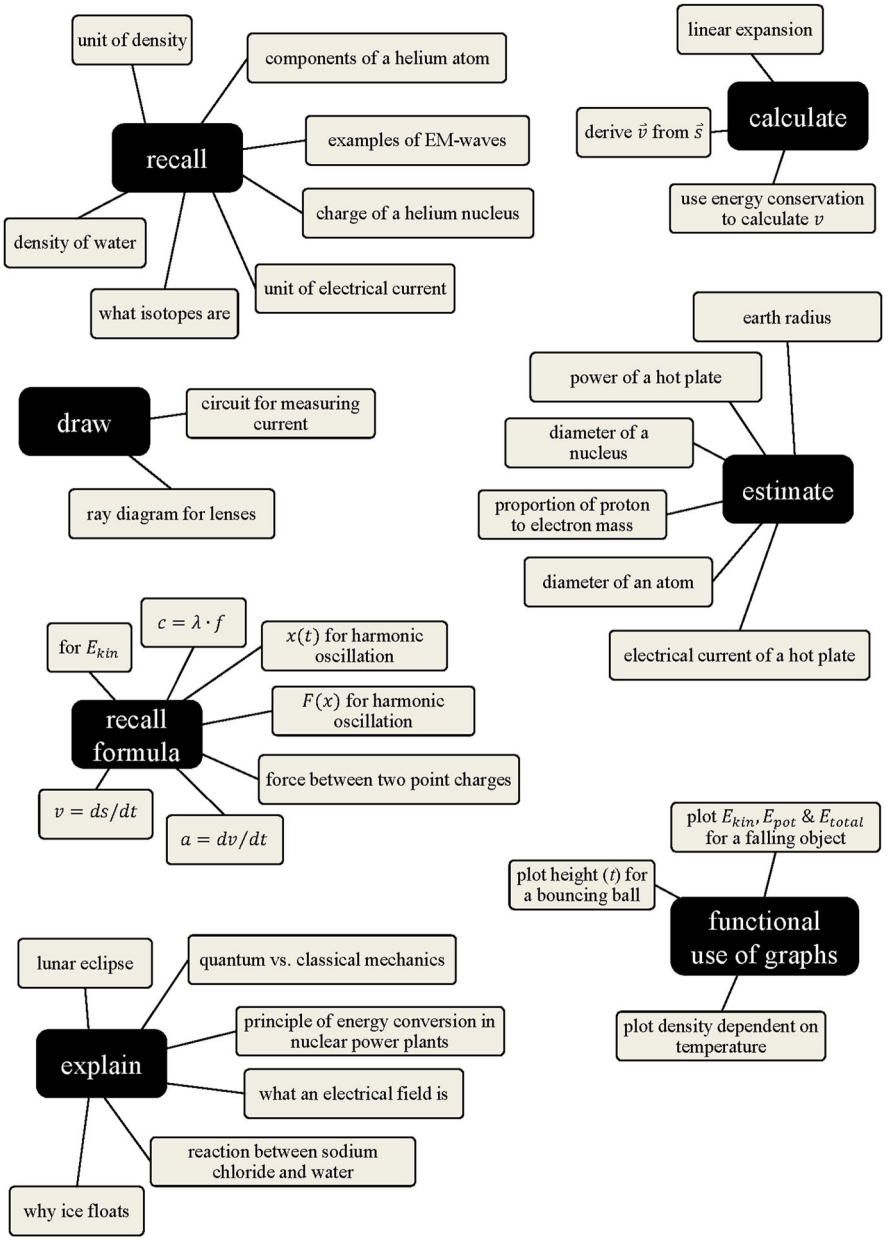

Figure 2. Description of physics items used in the test (Krause & Reiners-Logothetidou, 1981, pp. 311–316).

where students need to calculate or plot a relation between two variables for a specific situation. Here the student also needs to apply her/his knowledge in order to give a correct answer (quantity or graph). This is in line with the authors' statement that the test asked mainly for qualitative definition-related knowledge (Krause & Reiners-Logothetidou, 1981).

Empirical evidence for decreasing requirement-related knowledge or abilities

To our knowledge, there has not been any direct research based on empirical evidence regarding the differences in physics performance between different cohorts of physics freshmen over the past decades. The present study addresses this research gap. There was, however, some previous research on lower age groups related to our research question. Hodgen, Coe, Brown, and Küchemann (2016) tested 14-year-olds ($N \approx 7000$) in England in the areas algebra, decimals, ratio and fractions, and compared the results with those of an equivalent sample from 1976. The results showed that the overall understanding had decreased from 1976 to 2009.

In contrast to Hodgen et al.'s results, Kloosterman (2010) found significant gains in mathematics knowledge of students at the ages of 9, 13 and 17 years, in the U.S.A., from 1978 to 2004. However, the test instrument utilised by Kloostermann, the NAEP-LTT (The Long Term Trend National Assessment of Educational Progress), only covered a narrow range of procedural skills compared to the testing instrument in Hodgen et al.'s study (Hodgen et al., 2016).

Trends in International Mathematics and Science Study (TIMSS) Advanced (Mullis, Martin, Robitaille, & Foy, 2009) showed a significant decline from 1995 to 2008 in physics achievement in Norway and Sweden for students enrolled in physics courses in their final year of secondary school. This group typically included those students who afterwards pursued studies in science-related fields at university (Lie, Angell, & Rothagi, 2012). The results of students in the Russian Federation showed a statistically non-significant decline and Slovenia, a non-significant increase in physics achievement in TIMSS Advanced. In mathematics, there is evidence for a decline in Slovenia, Italy and Sweden and a minor increase for the Russian Federation. In both subjects, Sweden showed the most extreme decrease (more about this in the following section). Unfortunately, Germany did not participate in TIMSS Advanced 2008; and therefore, the results from none of the studies mentioned above can be generalised to the population of German freshmen.

The Cologne Institute for Economic Research (2014) described a thought experiment suggesting that mathematical competence of first semester university students has decreased in the last years. This thought experiment is a good example of one typical reasoning behind the statement 'more means worse'. It builds on the assumption that study places are assigned by ability; here the authors use PISA (Programme for International Student Assessment) mathematics scores of 2003 and 2009. If there are just minor changes in the ability distribution (here PISA score distribution), then higher enrolment automatically leads to lower mean scores for the group of freshmen.

However, it is unclear to what extent that assumption holds true. Additionally, physics freshmen differed systematically from the general cohorts of students in Germany (e.g. in

terms of mathematics knowledge or abilities). Therefore, it was not possible to infer from these considerations a decrease in the mathematical competence of the population of physics freshmen from 1978 to 2013.

Indirect evidence for a change in physics knowledge

As exemplified above, there is no clear empirical evidence that German physics freshmen nowadays demonstrate either a higher or lower average knowledge level in physics compared to the freshmen in 1978. However, in the following section, we list factors which were advantageous or disadvantageous for the students in the 2013 cohort.

If the premise 'more means worse' would hold true, then increasing freshmen enrolment rates could lead to a decline in the mean student knowledge. The thought experiment described above describes this logic.

Another argument includes changes in school standards or the way that physics is taught at the school level. Lie et al. (2012) argued that, among other factors, one factor leading to the decline in the performance of using mathematics in physics in TIMSS Advanced in Sweden could be the scientific literacy or the 'science for all' movement (opposed to a more specialist approach emphasising higher levels of mathematics in physics). In Germany, the PISA results published in 2001 accelerated the introduction of the term *competence* into the educational standards (Baumert et al., 2001; Klieme et al., 2003; Müller, Gartmeier, & Prenzel, 2013). In 2004, competency-based standards for lower secondary schools were introduced (Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany, 2004). These standards focus less on simple factual knowledge but emphasise conceptual understanding and cumulative learning (Müller et al., 2013). As the items of the entrance test from 1978 mainly required the reproduction of knowledge, the students in the 2013 cohort might underperform in these items due to these new educational goals. This is based on the assumption that curriculum changes influence teachers' behaviour sufficiently. Empirical studies show that this is not necessarily the case (Cronin-Jones, 1991; Vollstädt, 2003).

On the other hand, there are arguments supporting the assumption that the initial physics knowledge might have increased:

- Intensive courses in physics and mathematics: After 1978, Germany experienced a spread of intensive courses at school ('Leistungskurse'). These courses have more class time, cover more topics of the particular subject and foster a higher level of competency than basic courses. We assume a positive direct influence of the introduction of intensive courses in physics on the physics-related achievement of future physics students. Considering the importance of mathematical abilities for learning in physics, we further assume that the implementation of these courses in mathematics might positively influence physics-related learning gains indirectly.
- Preparatory courses: Before freshmen start their university studies in Germany, they often attend optional mathematics preparatory courses (Bausch et al., 2014). An online review of the supplied preparatory courses showed that the large majority of the preparatory courses teach mathematics and not physics (Buschhüter et al., 2016). As in the case of mathematics intensive courses, we assume that intensified mathematics knowledge or abilities might indirectly influence physics achievement.

On comparability between 1978 and 2013

The focus of this study was not to identify the causes for different prior knowledge of physics freshmen. Obviously, major transformations in the educational system, such as the competency-based school reform in Germany, may have caused differences in freshmen's knowledge, but there are several other influences to be considered. The characteristics of the freshmen cohort in 2013 were clearly different from those in 1978 in terms of gender, migration status and other individual variables of interest. We did not aim to achieve comparability of the two samples in terms of weighting or matching the two cohorts by these variables (e.g. by propensity score matching). Instead, we compared those two cohorts of freshmen who were given admission to physics studies in 1978 and 2013 in Germany to investigate the extent to which the freshmen in 2013 on average possessed a different level of physics knowledge.

Research question

The influences on student achievement of physics freshmen in 2013 presented above illustrate that it was not possible to form a hypothesis a priori on whether the students in the 1978 cohort sample outperform the 2013 cohort sample, or vice versa. Therefore, we derived the research question without a directional hypothesis: How do physics freshmen in 1978 and 2013 differ in their level of physics knowledge relevant for current physics programmes?

Research design and sample

To provide an answer to this research question, we used the complete test instrument previously applied in the 1978 assessment. All mathematics and physics items of 1978 were administered to the students in 2013. However, items were excluded from the data analysis if there was a need for a substantial change in the items. As the test information available from Krause and Reiners-Logothetidou (1981) included clear test instructions, it was possible to keep the test administration comparable to that in 1978. Before we explain the procedure of the study in more detail (see also Buschhüter et al., 2016), we report the pilot study, which is necessary in order to investigate the extent to which the instrument was usable.

Pilot study and modifications of the instrument

To evaluate if the test and the scoring rubric allow an objective assessment, we conducted a pilot study in the fall of 2012 with a sample of 159 physics students at a single university. Members of three different physics staff groups rated the item responses. The raters reported that some answers could not be distinctively coded as 'correct', 'false' or 'omitted'. Therefore, some items or related solutions in the scoring rubric were revised but then excluded from the analysis when biased item difficulties were anticipated (see the section on item exclusion process). In this way, these items could still be used to answer some other research questions (e.g. prediction of success in introductory physics); however, these questions are beyond the scope of this paper. For the same

reason, some of the self-report items (e.g. those on the students' perceived ability to conduct experiments) were replaced by other self-report questions. Due to the low cognitive load of these questions, we assumed that this change did not have an impact on the students' mathematics and physics performance in the test.

Testing procedure in 1978 and 2013

The testing procedure was carried out using the manual from 1978 (Krause & Reiners-Logothetidou, 1981). According to this procedure, the test time was limited to 90 min and the assessments were organised in the first two weeks of the freshmen's first semester at university. At one university in 2013, the test could only be scheduled for the beginning of the third week.

Sample

As in 1978, the sample in 2013 consisted of physics freshmen including students pursuing a teaching certification in physics. Table 1 shows the numbers of participants by study programmes in 1978 and 2013. In 2013, the sample also consisted of students who were studying in both programmes. These students were assigned to the group in the teaching certification programme. In 1978, the study was not conducted in the German Democratic Republic (Eastern Germany). Still, 39 Western German universities participated in 1978. The sample in 2013 consisted of 2322 students from 24 universities (including 4 Eastern Germany universities). We were able to obtain a data frame on the number of students who started studying physics in 2013 in Germany by university via the Federal Ministry of Statistics. Out of 45 physics departments in Germany, 39 participated in the 1978 assessment, which is equivalent to 87% of all the physics departments. In 2013, this rate was 40% (60 physics departments), including universities in the former German Democratic Republic. In 2013, 10 out of 16 Federal States participated. Therefore, we cannot argue that any of the two samples would be representative in a strong statistical sense, but the sample of 1978 covered a larger proportion of locations. We could assume the representativeness of this sample due to the large coverage of the population in 1978. At both times, the participation in this study was voluntary and initiated by the support of the conference of the physics faculties. The sample of universities in 2013 had a higher mean and median regarding the number of students enrolling in physics, $M = 228.88$, $SD = 183.35$ and $Mdn = 190$, compared to the population of all universities which offered a physics bachelor programme in 2013, $M = 178.86$, $SD = 155.07$ and $Mdn = 137$. However, the central tendencies (mean ranks) did not differ significantly in a two-sample Wilcoxon test (also known as the 'Mann–Whitney' test), $W = 844.5$,

Table 1. Number of students by study programme (Buschhüter et al., 2016, S. 66, with permission of Springer).

Programme	2013	1978[a]
Full Physics Bachelor	1959	1995
Teacher programme	363[b]	723
Total	2322	2718

[a]Krause and Reiners-Logothetidou (1981, p. 39).
[b]Some of these students studied in both programmes.

Table 2. Percentages of students with intensive courses ('Leistungskurse') at school (for 2013: based on 2250 students with German Abitur).

Intensive course	in Mathematics (%)		in Physics (%)	
	2013[a]	1978[b]	2013	1978[b]
Yes	54.6	~38	45.5	~35
No	36.2	~62	45.0	~65
N.A.	9.2	~0	9.5	~0

[a]Buschhüter et al. (2016, p. 66).
[b]Krause and Reiners-Logothetidou (1981, p. 97).

$p = .21$, $r = -0.13$. In this sense, we argue that the data analysis of this sample is informative in order to investigate tendencies in performance differences (calculations are based on the data of the Federal Ministry of Statistics).

In Germany, the *Gymnasium*[1] is traditionally the primary form of school to prepare students for university studies. In the sample of 2013, the proportion of students who finished their secondary education at a *Gymnasium* decreased by about 5% from 85.1% (Krause & Reiners-Logothetidou, 1981, pp. 90–91) to 80.1% (calculation based on 2250 students who had a German university entrance qualification). In contrast, the number of students participating in an intensive mathematics and physics course increased substantially (see Table 2). This is due to the fact that in 1978 intensive courses had just been introduced. In consequence, not all freshmen had the possibility to participate in such an intensive course yet (Krause & Reiners-Logothetidou, 1981). Additionally, about 70% of the students attended preparatory courses mainly covering mathematics preparation at university (Buschhüter et al., 2016). Preparatory courses were not so common in 1978. In the sample of 1978, only 27% of the students participated in these courses (Krause & Reiners-Logothetidou, 1981, p. 116).

Rating of student test item responses

For the majority of test items, the scoring categories were 'correct response', 'incorrect response' and 'omitted response'. For some items, the scoring rubric demanded additional coding ('incomplete response' for 3 items; 'partially correct response' for 1 item and 'calculation error' for 1 item). As in 1978, the local staff at the participating universities carried out the primary rating. At two universities, the booklets were transferred to and rated at the University of Potsdam.

To investigate the inter-rater agreement, we conducted a second rating of the student item responses. We drew a probability sample for each university, which included all local participating students. All samples consisted of a minimum of ten booklets or at least 10% of the local sample size.

Two raters carried out the secondary rating. This rating was conducted item-wise in order to avoid that an inter-rater agreement between these three raters had to be investigated additionally. Before investigating the inter-rater agreement, the answers were scored as '1' for 'correct' and '0' for 'not correct' including all other responses ('incorrect', 'partially correct', etc.).

For the first rating, it was not known how student responses were assigned to the raters. When interpreting coefficients of inter-rater agreement, it should be noted that the first and second raters were indeed 'pseudo-raters' (i.e. ratings from several persons were

merged into the ratings of one single rater as described above). We deliberately decided not to use a multifaceted design (Boone, Staver, & Yale, 2014) as the data had to be analysed using classical test theory.

Method of analysis

As the data set of 1978 was not available, the different item difficulties reported by Krause and Reiners-Logothetidou (1981, pp. 305–316) were used as reference. Krause and Reiners-Logothetidou also reported distributions of person raw scores (p. 154). These distributions could not be used for the comparison because some items had to be excluded from the analysis. The variance of the reported raw scores after exclusion of these items was unknown. As the data frame was a complete matrix with dichotomous values (1, 0), the new mean person scores could be derived from averaging the item difficulties; it is identical to the difficulty of the test (the mean item difficulty).

To conduct a significance test and calculate an effect size, we defined the standard deviation of raw scores in 1978 (in percentage) to be identical with the standard deviation of 2013 ($SD_{2013} = 19.9\%$). This assumption is reasonable because the reported standard deviation (without the exclusion of items) of 1978 is almost identical ($SD_{1978,\ \text{reported}} = 20.9\%$; Krause & Reiners-Logothetidou, 1981, p. 154)

Additionally, we assumed that the rater severity did not change on average. The low inferential guidance of the scoring rubric, the level of inter-rater agreement (see the section on inter-rater agreement) and the high number of raters involved in the scoring process supported this assumption.

Results: item exclusion process and test quality

Item exclusion process

Before presenting the results of the comparison between the freshmen in 1978 and 2013, we need to ensure that person measures in this test are reliable and valid. The physics test originally consisted of 47 items. It was necessary to exclude 13 items due to various reasons. We summarise the item exclusion process in the three steps below:

(1) The item was not valid in terms of content-related requirement: Four interviews were conducted with experts (see the section on validity) to determine whether an item should be excluded from the item pool. One item was additionally excluded because it was clearly outdated, asking for ways to amplify electrical current in a radio (altogether five items were thus excluded).
(2) Insufficient inter-rater agreement: From Cohen's κ, a modified coefficient κ^* (Equation (1)) was developed:

$$\kappa^* = \frac{\overline{P_0} - \overline{P_c}}{1 - \overline{P_c}}, \tag{1}$$

where $\overline{P_0}$ is the weighted mean of the agreements of the pseudo-raters and $\overline{P_c}$ is the equivalent for the agreements by chance. The weights are the numbers of physics freshmen at the local university. If university staff A rated more booklets than staff group B, then the behaviour of A influenced the item difficulty more than B did.

This is why the rater behaviour related to larger samples needed to be weighted in a stronger way compared to the rater behaviour of universities contributing a smaller number of freshmen to the main sample.

When the coefficient fell below a critical value of $\kappa^* = 0.6$ (Wirtz & Caspar, 2002, p. 59), the item was excluded from the analyses (five additional items were thus excluded). Problems in inter-rater agreement occurred, for instance, on an item demanding students to plot a graph or a question asking them to explain the interference of electromagnetic waves using a sketch. These items allowed a more subjective interpretation of correctness.

(3) The item was not comparable with the same item administered in 1978: In the pilot study, systematic difficulties in the rating process according to the scoring rubric from 1978 were uncovered. Thus, we changed some of the items and/or the scoring rubric for these items. When changes in item difficulties were expected in an item, we excluded that item from the analysis (three additional items were thus excluded).

The resulting item pool consisted of 34 (=47 − 5 − 5 − 3) items for the following analysis.

Inter-rater agreement

Appendix 1 shows the histogram of κ^* for the remaining items. The mean and median κ^* reached a value of 0.85, indicating a very acceptable inter-rater agreement (Wirtz & Caspar, 2002, p. 59). Falling into the interval of 0.60–0.75, even the minimum value of 0.66 may still be considered as satisfactory according to common standards (Wirtz & Caspar, 2002, p. 59).

Reliability and discrimination

We found a high level of test reliability (Cronbach's $\alpha = 0.88$), although a few items demonstrated hardly sufficient item discrimination values (see Appendix 2). As these item discrimination values were still clearly positive, and in order not to restrict the test content, these items remained in the item pool to be analysed.

Validity

To ensure the items covered relevant physics content, four content matter experts reviewed the items. German physics lectures can traditionally be separated into courses for experimental and theoretical physics. These lectures are usually taught by professors and accompanied by tutorial groups. Thus, we chose one professor and one tutor teaching experimental physics and one professor and one tutor teaching theoretical physics to review the instrument. An item was removed from the item pool when at least two experts from the different perspectives (experimental or theoretical) agreed that the knowledge required to solve that item was neither necessary nor helpful for students before the end of the undergraduate studies. Additionally, at least one of the two experts needed to provide a valid justification.

For five universities, we assessed the power of raw scores to predict university grades for bachelor students after the first semester by calculating the Pearson correlation coefficients

Table 3. Correlations between raw scores and university grades after the first semester for five different universities.

University	Pearson's *r*	*N*	Type of grade
1	−0.20	14	Average
2	−0.22	19	Average
3	−0.39	124	Average
4	−0.80	31	Physics 1
5	−0.35	53	Physics 1

between both measures (see Table 3). Students pursuing a teaching certification were excluded due to the low sample size. The correlations were computed separately for each university in order to account for the different kinds of grades available. Three universities submitted average student grades and two universities submitted student grades of the first semester physics exam ('Physics I'). The average grades consisted mostly of physics and mathematics grades (for further information about the grades, see Buschhüter et al., 2016).

All correlations were expected to be negative as the German grading system includes grades from 5 (lowest grade) to 1 (highest grade). The missing values were excluded from the calculations, which might possibly lead to an underestimation of the correlations. Using Fisher *z*-transformation, we calculated the mean correlation ($r = -0.43$). Summarising these results on the validity of the test scores, we found that the correlations showed evidence of predictive power in all five samples. The correlations for university 1 ($N = 14$) and 2 ($N = 19$) were considered as low.

Missing values and the test length effect

As in the 1978 assessment, omitted items were coded as 'incorrect response' in order to achieve comparable results. For the same reason, we decided not to exclude any items from the instrument itself and maintain the same testing time. Appendix 3 graphically shows the impact of the test length effect, displaying the percentage of responses coded as 'not reached' for each item. The results indicated that the test was actually speeded and underlined the necessity to administer the test with the same testing time and number of items.

However, we had strong evidence to believe that the test length effect of 1978 and 2013 had a similar impact. Krause and Reiners-Logothetidou (1981) reported that 14% of the physics freshmen in 1978 did not answer the items on the last two pages (p. 37). This percentage was identical to that in 2013, which indicated that the test length effect was similar for both cohorts.

Results: comparison of physics knowledge

Comparison of student test scores

A two-tailed independent *t*-test showed that the mean of the distribution of 1978 (M_{1978} = 44.62%) was significantly higher than the mean of 2013, M_{2013} = 34.03%, $t\ (5038) = 18.83$, $p < .001$. Due to equal standard deviations (see the section on method of analysis), Cohen's

d with pooled standard deviation simplifies to Equation (2):

$$d = \frac{M_{1978} - M_{2013}}{SD_{2013}}. \tag{2}$$

In this case, the effect size is $d = 0.53$, which is considered a medium effect (Cohen, 1988).

Comparison on the item level

Appendix 4 shows the item difficulties in 1978 and 2013. Obviously, almost all the items displayed higher item difficulties in 2013 compared to the results from 1978.

We calculated the differences Δp of item difficulties as:

$$\Delta p = p_{2013} - p_{1978}, \tag{3}$$

with p_{2013} as the item difficulties from 2013 and p_{1978} as those from 1978. Analogous to Figure 1, we displayed these differences separately for each content area of physics (see Figure 3(a)).

It needs to be stressed that the test was not developed to include consistent and highly reliable subscales. Thus, some content areas are represented by a single item. Table 4 summarises the knowledge considered to be necessary to solve the items correctly (alongside the item difficulty levels, and the related differences in item difficulty).

Another way to categorise and characterise the items is to group them by the activities that are usually necessary to give the correct responses (Figure 3(b)).

The categories of activities were developed inductively (Mayring, 2014). In an expert rating, we assigned the items again to the same categories and found very acceptable values of inter-rater agreement (Cohen's $\kappa = 0.79$), with an agreement of 82.4% ($N =$ 34 items). The activities are described below:

- *Estimate:* To solve these items, students needed to estimate or recall values of physical quantities from any kind of content area (length, relation of masses, electric current and power).
- *Draw:* These items required drawings that are physics specific, but not graphs. Two items represented the category: a ray diagram and a circuit.
- *Recall Formula:* This category consisted of items where students needed to recall a formula. Except for P40, all of the items were part of the content area of mechanics.
- *Explain:* Students were requested to explain or describe a phenomenon with items classified to this category. The explanations required always contain some amount of the reproduction of knowledge.
- *Recall fact(s):* To master these items, students simply needed to recall one or more facts.
- *Calculate:* These items required calculations or other mathematical operations.
- *Functional use of graphs:* Items from this category required using a graph to describe the functional relations between physical quantities.

Figure 3(b) displays the differences in item difficulties depending on the types of activities. This plot shows that only items requiring plotting a graph did not differ in their mean item difficulty. What distinguishes this activity from the other ones? Except for

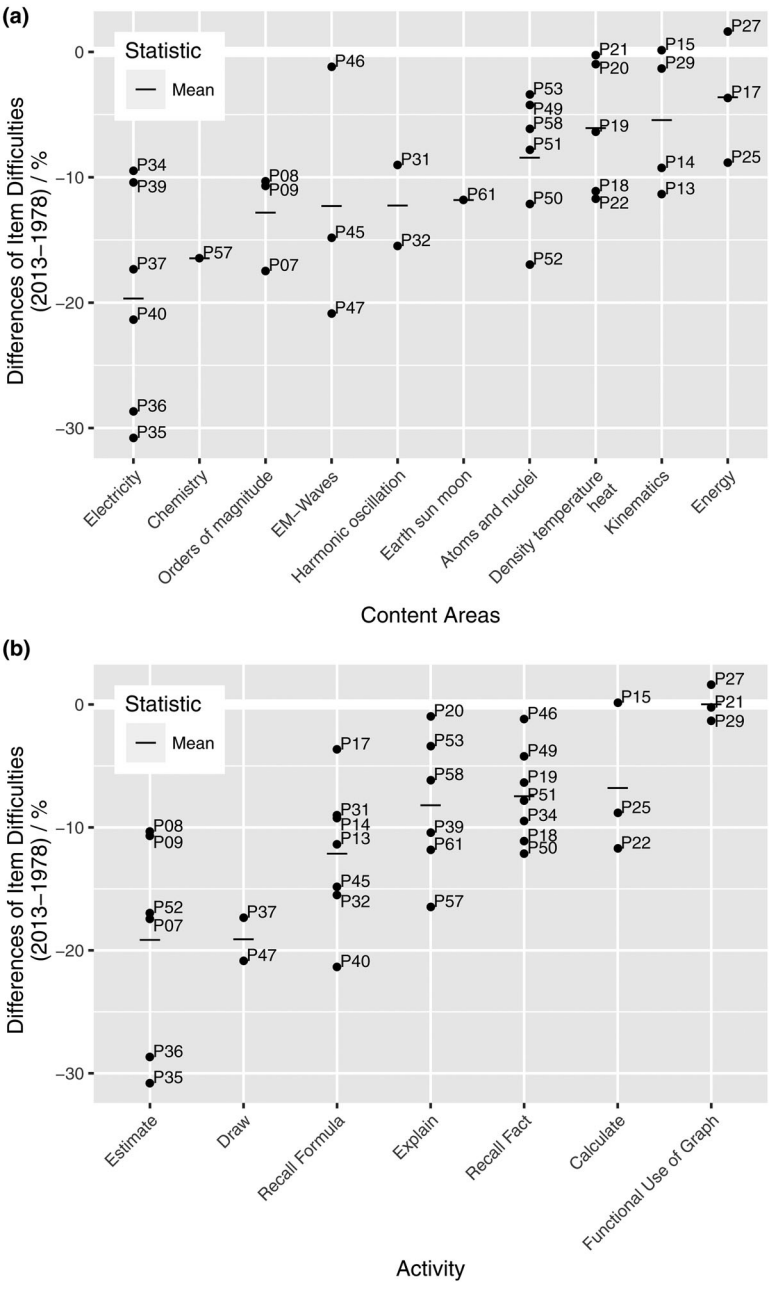

Figure 3. Differences in item difficulties in 1978 (Krause & Reiners-Logothetidou, 1981, pp. 311–316) vs. 2013 separated by content area (a) and by underlying activity (b).

the activity 'Calculate', all the other areas can be regarded as mainly requiring the reproduction of knowledge. Even if these items could not be judged as completely independent from prior knowledge (e.g. P27), they mainly required an intuitive understanding of the problem. For instance, plotting the height–time function of a bouncing ball mainly relied on the ability to display the relation between two quantities and less on

Table 4. Item description, difficulties and their differences for 2013 and 1978 (Krause & Reiners-Logothetidou, 1981, pp. 311–316) by content area.

Content area	Item	Description (student can …)	p_{1978} (%)	p_{2013} (%)	Δp (%)
Electricity	P34	*recall* unit of electrical current	86.6	77.1	−9.5
	P35	*give estimate* for electrical current hot plate	44.4	13.6	−30.8
	P36	*give estimate* for power of hot plate	44.4	15.7	−28.7
	P37	*draw* circuit for measuring current	53.0	35.7	−17.3
	P39	*explain* what an electrical field is	36.0	25.6	−10.4
	P40	*recall* formula of force between two point charges	37.2	15.8	−21.4
Chemistry	P57	*explain* reaction between sodium chloride and water	43.9	27.4	−16.5
Orders of magnitude (length)	P07	*give an estimate* of the earth radius	58.5	41.0	−17.5
	P08	*give an estimate* of the diameter of an atom	32.3	22.0	−10.3
	P09	*give an estimate* of the diameter of a nucleus	26.1	15.4	−10.7
EM-Waves	P45	*recall formula*: $c = \lambda \cdot f$	43.6	28.8	−14.8
	P46	*recall* examples of EM-waves	53.3	52.1	−1.2
	P47	*draw* ray diagram for lenses	36.8	15.9	−20.9
Harmonic oscillation	P31	*recall formula:* $x(t)$ for harmonic oscillation	28.1	19.1	−9.0
	P32	*recall formula:* $F(x)$ for harmonic oscillation	31.9	16.4	−15.5
Earth, sun, moon	P61	*explain* lunar eclipse	63.5	51.7	−11.8
Atoms and nuclei	P49	*recall* components of a helium atom	49.4	45.2	−4.2
	P50	*recall* charge of helium nucleus	52.6	40.5	−12.1
	P51	*recall* what isotopes are	49.4	41.6	−7.8
	P52	*estimate* the proportion of proton to electron mass	31.0	14.0	−17
	P53	*recall the* principle of energy conversion in nuclear power plants	41.5	38.1	−3.4
	P58	*explain* difference: quantum vs. classical mechanics	12.0	5.9	−6.1
Density, temperature, heat	P18	*recall* unit of density	70.2	59.1	−11.1
	P19	*recall* density of water	48.9	42.5	−6.4
	P20	*explain* why ice floats	80.9	79.9	−1.0
	P21	*plot* density dependent on temperature	30.4	30.1	−0.3
	P22	*calculate* linear expansion	43.2	31.5	−11.7
Kinematics	P13	*recall formula*: $v = ds/dt$	44.6	33.2	−11.4
	P14	*recall formula* : $a = dv/dt$	42.8	33.5	−9.3
	P15	*derive* $\vec{v}$ from $\vec{s}$	17.5	17.7	0.2
	P29	*plot* height (t) for bouncing ball	52.4	51.1	−1.3
Energy	P17	*recall formula* for E_{kin}	67.4	63.7	−3.7
	P25	use energy conservation to *calculate* v	46.6	37.8	−8.8
	P27	*plot* E_{kin}, E_{pot} and E_{total} for a falling object	16.6	18.2	1.6

declarative knowledge (as everyone has once seen a bouncing ball). Plotting the function for the density of air depending on the temperature required minimal and basic declarative knowledge (the higher the temperature, the lower the density or air expands when it is heated). We found that the items requiring application showed less negative mean differences (between 2013 and 1978) than did those mainly requiring the reproduction of knowledge. Items with the most extreme negative differences strongly depended on the reproduction of knowledge (e.g. item P35: 'How high is the current in an electrical hot plate?' and item P40: 'What is the law for the force between two electrical point charges?'), whereas item P15 in contrast relied on the application of rather fundamental knowledge 'The trajectory of point mass is described as $\vec{r}(t) = 3\vec{g}t^2 - \vec{v}_0 t + \vec{r}_0$. Derive the velocity of the point mass at the time t'. The latter displayed a difference in item difficulties of about zero. The items within quotation marks were translated by the authors.

It needs to be emphasised that the relation between differences in item difficulties and the activities or the content areas is correlational, as items were not designed to differ solely in these properties.

Discussion

The aim of this study was to investigate the extent to which physics freshmen in the cohort in 2013 in Germany differed in their physics knowledge from their counterparts in 1978. The results provide evidence that the physics freshmen in 2013 possessed a lower average knowledge level compared to the freshmen in 1978, and the difference was statistically significant and of medium effect size. Analysing the difference in more detail showed that this result was surprisingly unambiguous and the evidence, therefore, has to be regarded as strong: Almost all items were solved proportionately less frequently in 2013 than in 1978.

However, there was considerable variance within the differences in item difficulties, reaching from −30.8% to 1.6%. We argue that the physics knowledge assessed by the test was mainly concerned with the reproduction of knowledge. Items involving plotting relations between two physical quantities showed no differences in their mean item difficulties.

Sampling-related limitations of this study include that participation in the study was voluntary for the universities and that the sample in 1978 consisted of only West German universities due to the political separation between East and West Germany until 1989. Furthermore, we need to assume equal rater severity and equal variances in person scores. The latter also means that we cannot compare the variances of the two samples to investigate the extent to which the negative difference in mean performance would be caused, for example, by a tail of lower values in the person score distribution. As mentioned in the beginning of this paper, we cannot conclude from the changes in the test performance that increasing freshmen enrolment rates were responsible for the observed differences in physics knowledge in these two cohorts of physics freshmen. We can, however, investigate the changes in the prerequisite physics knowledge in a system undergoing educational expansion. In the following section, we discuss the implications of this study.

Implications

Communication between teachers and students

We have discussed that physics knowledge at the university does not develop independently from prior physics knowledge. University physics teachers in Germany need to be aware that the declarative knowledge base of physics freshmen nowadays is on average probably weaker than it was in 1978. Krause and Reiners-Logothetidou (1981) mentioned that the test was aimed to measure whether the students could recall common elements of the communication between university teachers and students. These elements seem less available to the freshmen in 2013 than to those in 1978; therefore, university teachers should be careful to rely on such elements. As this study shows that with time, knowledge changes, we emphasised the need for regularly assessing the initial knowledge state of freshmen (see a similar conclusion in Hodgen et al., 2016).

Perceptions of university teachers

As mentioned in the introduction of the paper, there is empirical evidence to believe that many university teachers think that students nowadays are not as talented as were the

students in the past. The present study supports this kind of assumption of university teachers and indicates that we should take these worries about the decrease in knowledge of freshmen students seriously. The mathematics test results (Buschhüter et al., 2016), however, also showed that we should not generalise from these results. Mathematics knowledge of freshmen did not show the same overall decline, but its increases and declines strongly depended on the actual demands of an item in the test. Therefore, caution should be taken against imprudent generalisations, across disciplines, traits, down to the level of a single test item.

Can educational expansion be implemented without a decrease in mean student knowledge?

The students in the 2013 cohort did not underperform compared to the students in the 1978 cohort in the mathematics test, but they did underperform in the physics test. In both cohorts, the students answering the mathematics test were the same individuals who answered the physics test, while the different constructs (physics and mathematics knowledge) had undergone different treatments prior to the students taking the test (preparation at school, preparatory courses at university, etc.). This supports the notion that the performance resources are sensitive to the treatment provided to the students before they begin their studies. It also supports the notion of malleability of these resources and that 'MORE' does not necessarily have to mean 'WORSE', if we provide treatment which is well-aligned with the requirements of the university programmes.

Note

1. The 'Gymnasium' is the primary traditional school form in Germany to prepare students for university studies (for a summary of the German educational system, see Bonsen, Bos, & Frey, 2008).

Acknowledgements

We hereby wish to thank all participating members of the university staff for organising the test and coding the student answers and all participating students for taking the test. We would also like to thank the Wilhelm and Else Heraeus Foundation for funding this study and the German conference of the physics faculties for its general support. Special thanks go to the physics staff groups in Aachen, who participated and supported the pilot project.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Wilhelm and Else Heraeus Foundation.

References

Albrecht, A. (2011). *Längsschnittstudie zur Identifikation von Risikofaktoren für einen erfolgreichen Studieneinstieg in das Fach Physik* [A logitudinal study to identify risk factors for a successfull start in physics programs] (Doctoral Dissertation, Free University of Berlin, Berlin, Germany). Retrieved from http://www.diss.fu-berlin.de/diss/servlets/MCRFileNodeServlet/FUDISS_derivate_000000010456/Dissertation_Druckversion_Andre_Albrecht_UB.pdf

Amis, K. (1960, July 6–11). Lone voices. *Encounter*, *XV*. Retrieved from https://www.unz.org/Pub/Encounter-1960jul-00006?View=PDF

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., … Weiß, M. (Eds.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* [An international comparison of school students basic competences]. Opladen: Leske + Budrich.

Bausch, I., Biehler, R., Bruder, R., Fischer, P. R., Hochmuth, R., Koepf, W., … Wassong, T. (Eds.). (2014). *Mathematische Vor- und Brückenkurse: Konzepte, Probleme und Perspektiven* [Mathematics preperatory courses: Concepts, problems and perspectives]. Wiesbaden: Springer Spektrum.

Bonsen, M., Bos, W., & Frey, K. A. (2008). Germany. In I. V. Mullis, M. O. Martin, J. F. Olson, D. R. Berger, D. Milne, & G. M. Stanco (Eds.), *TIMSS 2007 encyclopedia* (Vol. 1, pp. 203–216). Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer.

Buschhüter, D., Spoden, C., & Borowski, A. (2016). Mathematische Kenntnisse und Fähigkeiten von Physikstudierenden zu Studienbeginn [Mathematics knowledge of first semester physics students]. *Zeitschrift für Didaktik der Naturwissenschaften*, *22*(1), 61–75. doi:10.1007/s40573-016-0041-4

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.

Cologne Institute for Economic Research. (2014). *Bildungsmonitor 2014: Die richtigen Prioritäten setzen [Educational monitoring 2014: Setting the right priorities]*. Retrieved from https://www.iwkoeln.de/_storage/asset/179418/storage/master/file/5009239/download/Bildungsmonitor%202014%20mit%20Ergebnisbericht_18082014.pdf

Cronin-Jones, L. L. (1991). Science teacher beliefs and their influence on curriculum implementation: Two case studies. *Journal of Research in Science Teaching*, *28*(3), 235–250. doi:10.1002/tea.3660280305

Driscoll, M. P. (2005). *Psychology of learning for instruction*. Toronto: Allyn and Bacon.

Duke, C. (2005). Values, discourse and politics: An Australian comparative perspective. In T. Tapper & D. Palfreyman (Eds.), *Understanding mass higher education* (pp. 28–50). London: RoutledgeFalmer.

Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics*, *53*(11), 1043–1048. doi:10.1119/1.14030

Hazari, Z., Tai, R. H., & Sadler, P. M. (2007). Gender differences in introductory university physics performance: The influence of high school physics preparation and affective factors. *Science Education*, *91*(6), 847–876. doi:10.1002/sce.20223

Hodgen, J., Coe, R., Brown, M., & Küchemann, D. (2016). *Educational standards over time: Changes in mathematical understanding between 1976 and 2009 in England*. Manuscript submitted for publication.

Hudson, H. T., & McIntire, W. R. (1977). Correlation between mathematical skills and success in physics. *American Journal of Physics*, *45*(5), 470–471. doi:10.1119/1.10823

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., … Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise* [About the development of national educational standards. An expertise]. Berlin: Federal Ministry of Education and Research.

Kloosterman, P. (2010). Mathematics skills of 17-year-olds in the United States: 1978 to 2004. *Journal for Research in Mathematics Education*, *41*(1), 20–51. Retrieved from http://www.jstor.org/stable/40539363

Kragh, H. (2015). Mathematics and physics: The idea of a pre-established harmony. *Science & Education*, *24*(5), 515–527. doi:10.1007/s11191-014-9724-8

Krause, F., & Reiners-Logothetidou, A. (1981). *Kenntnisse und Fähigkeiten naturwissenschaftlich orientierter Studienanfänger in Physik und Mathematik: Die Ergebnisse des bundesweiten Studieneingangstests Physik 1978* [Knowledge and abilities of first year students in science-related fields regarding physics and mathematics: The results of the national students entrance test 1978]. Bonn: University of Bonn.

Lie, S., Angell, C., & Rothagi, A. (2012). Interpreting the Norwegian and Swedish trend data for physics in the TIMSS Advanced study. *Nordic Studies in Education*, *32*, 177–195.

Long, D. D., McLaughlin, G. W., & Bloom, A. M. (1986). The influence of physics laboratories on student performance in a lecture course. *American Journal of Physics*, *54*(2), 122. doi:10.1119/1.14705

Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. misc. Retrieved from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173

Müller, K., Gartmeier, M., & Prenzel, M. (2013). Kompetenzorientierter Unterricht im Kontext nationaler Bildungsstandards [Competence oriented teaching in the context of the national educational standards]. *Bildung und Erziehung*, *66*(2), 127–144. doi:10.7788/bue.2013.66.2.127

Mullis, I. V. S., Martin, M. O., Robitaille, D. F., & Foy, P. (2009). *TIMSS Advanced 2008. International report. Findings from IEA's study of achievement in advanced mathematics and physics in the final year of secondary School*. Retrieved from http://pirls.bc.edu/timss_advanced/downloads/TA08_International_Report.pdf

Sadler, P. M., & Tai, R. H. (2001). Success in introductory college physics: The role of high school preparation. *Science Education*, *85*(2), 111–136.

Shumba, O., & Glass, L. W. (1994). Perceptions of coordinators of college freshman chemistry regarding selected goals and outcomes of high school chemistry. *Journal of Research in Science Teaching*, *31*(4), 381–392.

Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany [SCMECA]. (2004). *Bildungsstandards der Kultusministerkonferenz: Erläuterungen zur Konzeption und Entwicklung* [Educational standards of the SCMECA: Explanations about their conceptions and their developement]. München: Luchterhand. Retrieved from http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Konzeption-Entwicklung.pdf

Tapper, T., & Palfreyman, D. (2005a). Conclusion: The reshaping of mass higher education. In T. Tapper & D. Palfreyman (Eds.), *Understanding mass higher education* (pp. 247–261). London: RoutledgeFalmer.

Tapper, T., & Palfreyman, D. (Eds.). (2005b). *Understanding mass higher education: Comparative perspectives on access*. London: RoutledgeFalmer.

Tuminaro, J. (2004). *A cognitive framework for analyzing and describing introductory students' use and understanding of mathematics in physics* (Doctoral dissertation). Retrieved from http://www.physics.umd.edu/perg/dissertations/Tuminaro/TuminaroPhD.pdf

Uzuntiryaki, E., Boz, Y., Kirbulut, D., & Bektas, O. (2010). Do pre-service chemistry teachers reflect their beliefs about constructivism in their teaching practices? *Research in Science Education*, *40*(3), 403–424. doi:10.1007/s11165-009-9127-z

Vollstädt, W. (2003). Steuerung von Schulentwicklung und Unterrichtsqualität durch staatliche Lehrpläne? [Management of school developement and quality of teaching by state curricula?] *Zeitschrift für Pädagogik*, *47*(Suppl.), 194–214.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen* [Inter-rater agreement and inter-rater reliability:

Methods to determine and improve the reliability of ratings using systems of catgories and rating scales]. Göttingen: Hogrefe.
Yorke, M., & Knight, P. (2004). Self-theories: Some implications for teaching and learning in higher education. *Studies in Higher Education*, *29*(1), 25–37.
Young, H. D., Freedman, R. A., Ford, A. L., & Sears, F. W. (2012). *Sears and Zemansky's university physics: With modern physics* (13th ed.). San Francisco: Pearson Addison-Wesley.

Appendix 1

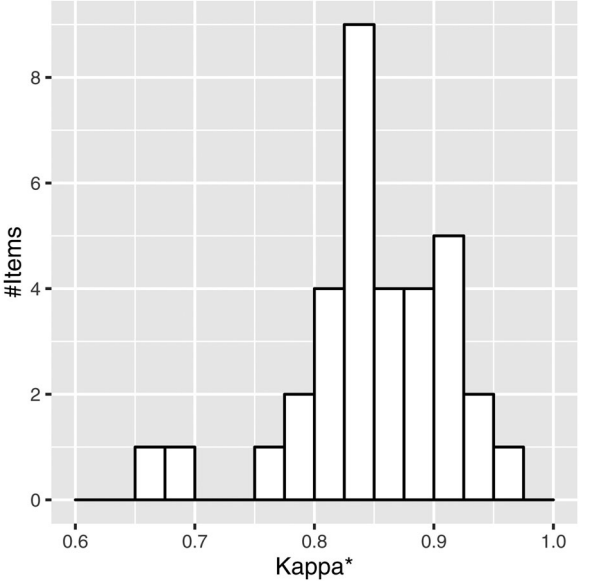

Figure A1. Histogram of coefficients κ^* for items.

Appendix 2

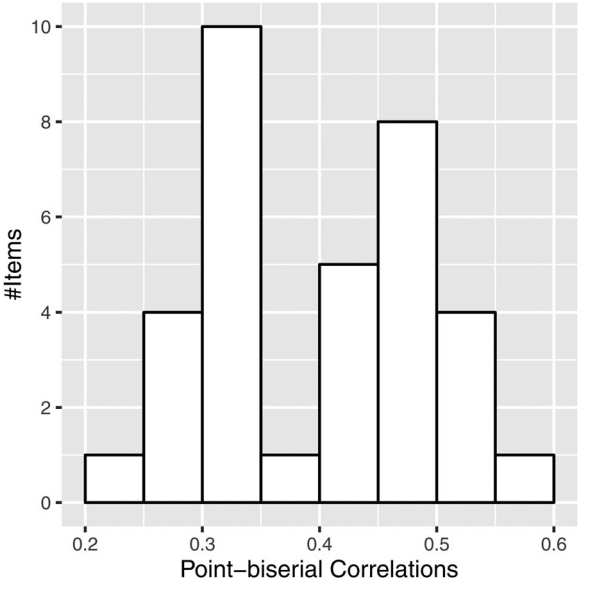

Figure A2. Histogram of point-biserial correlations between items score (0, 1) and total score excluding the item of focus.

Appendix 3

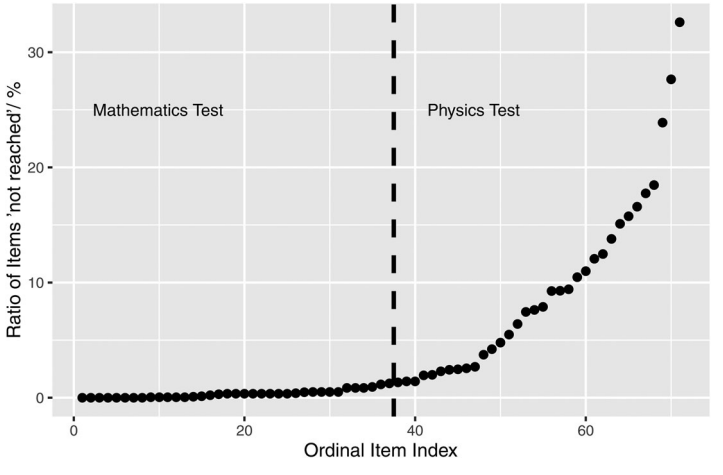

Figure A3. Items coded as 'not reached' by item order.

Appendix 4

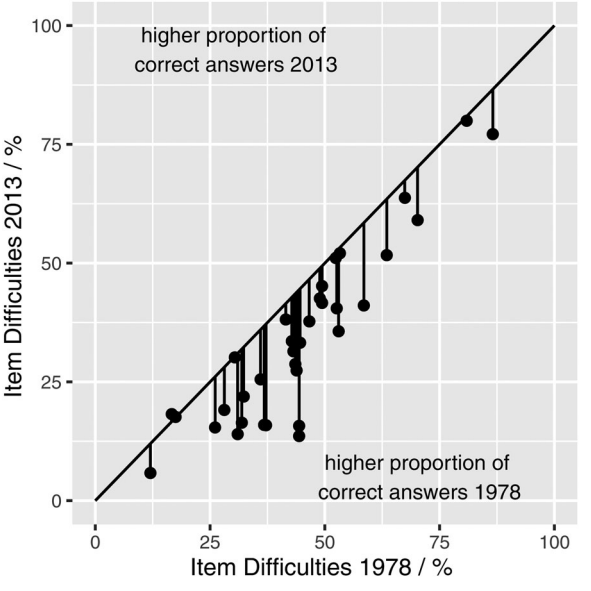

Figure A4. Scatterplot of item difficulties in 1978 (Krause & Reiners-Logothetidou, 1981, pp. 311–316) vs. 2013.