Large-scale Assessments in Education

SHORT REPORT

Open Access

The standard setting process: validating interpretations of stakeholders

Nele Kampa*, Helene Wagner and Olaf Köller

*Correspondence:
kampa@ipn.uni-kiel.de
Leibniz Institute for Science and Mathematics Education at the Christian-Albrechts-University of Kiel, Olshausenstrasse 62, 24118 Kiel, Germany

Abstract

Background: Stakeholders' interpretations of the findings of large-scale educational assessments can influence important decisions. In the context of educational assessment, standard-setting remains an especially critical element, because it is complex and largely unstandardized. Instruments established by means of standard-setting procedures such as proficiency levels (PL) therefore appear to be arbitrary to some degree. Owing to the significance such results take on, when they are communicated to stakeholders or the public, a thorough validation of this process seems crucial. In our study, ministry stakeholders intended to use PL established in an assessment of science abilities to obtain information about students' strengths and weaknesses regarding science abilities in general and specifically about the extent to which they were prepared for future science studies. The aim of our study was to investigate the validity arguments regarding these two intended interpretations.

Methods: Based on a university science test administered to 3641 upper secondary students (Grade 13), a panel of nine experts set four cut scores using two variations of the Angoff method, the Yes/No Angoff method (multiple choice items) and the extended Angoff method (complex multiple choice items). We carried out t-tests, repeated measures ANOVA, G-studies and regression analyses to support the procedural, internal, external, and consequential validity elements regarding the aforementioned interpretations of the cut scores.

Results: Our t-tests and G-studies showed that the intended use of the cut scores was valid regarding procedural and internal aspects of validity. These findings were called into question by the experts' lack of confidence in the established cut scores. Regression analyses including number of lessons taught and intended and pursued science-related studies showed good external and poor consequential validity.

Conclusion: The cut scores can be used as an indicator of 13th graders' strengths and weaknesses in science. They should not be used as an indicator for preparedness for science university studies. Since assessment formats are continually evolving and consequently leading to more complex designs, further research needs to be conducted on the application of new standard-setting methods to meet the challenges arising from this development.

Keywords: Standard setting, Validity, Science education, Extended Angoff method, Yes/No Angoff method, Large-scale assessment

Introduction

Most large-scale assessments (LSA) aim to provide findings on the effectiveness of a school system to different stakeholders (e.g., ministry personnel, teachers or schools). Within these accountability programs, results are almost always reported in a standard-based way (Haertel 2002). Proficiency level models (PLM) map raw continuous scores into performance categories. The tests, methods, and analyses for these accountability programs are becoming more and more complex, which poses new challenges for standard-setting procedures. This article describes a validity study that investigates validity arguments for the interpretation of cut scores derived from a standard-setting procedure applied to a science ability scale of 13th graders. As the items for this scale contained multiple item formats, the reader gets insights into one example of new challenges that evolving and more complex assessments entail for standard-setting procedures.

The study was commissioned by the Ministry of School and Professional Education of a German federal state with the aim of providing information on (a) the students' strengths and weaknesses in science when they graduated from upper secondary school in that federal state, and (b) the extent to which these students were prepared for university studies in science. Thus, the main purpose of the assessment was a monitoring function. The results of the assessment were presented to stakeholders in education and the published book (Leucht et al. 2016) was distributed to all upper secondary schools in this federal state.

Germany does not have proficiency levels (PL) for the end of upper secondary education (Grade 13) and a mere scale (i.e., the average performance of the students) would not be sufficient to support the stakeholders' intended interpretations of the results. Therefore, standard-setting experts developed PL as well as proficiency level descriptions (PLD) during a standard-setting process which integrated the different item formats. As the Angoff method has been adapted to various testing conditions and showed equivalency or superiority to other test-centered standard-setting methods such as the Ebel method (e.g., Ebel 1972; Yudkowski et al. 2008), we applied two variations of the Angoff method—the Yes/No method (multiple choice [MC] items; Impara and Plake 1997) and the extended Angoff method (complex multiple choice [CMC] items; Cizek and Bunch 2007). In our study, we provide validity arguments for the two interpretations of the PLM proposed by the ministry stakeholders on the basis of cut scores derived by applying these two Angoff methods.

Not many validity studies associated with standard-setting procedures have been carried out and most of them focus on procedural and internal validity arguments (e.g., Freunberger 2013; Hurtz and Auerbach 2003; McGinty 2005). Over the last decades, the focus of validity theory has changed as regards the interpretation and usages of assessments and results of standard-setting procedures (Kane 2013). In consequence, the validation of the interpretation and the use of cut scores derived from these standard-setting procedures has to change accordingly. In our study, we aim to represent this change.

The concept of validity

The concept of validity underwent major changes. Messick (1995) proposed to unify different validity conceptualizations (e.g., construct validity, incremental validity). This approach led to an integrated definition of validity focusing on the intended and unintended interpretations of test scores (Kane 2013). Presently, a prominent definition of validity conceptualizes it as an evaluation of "the plausibility of the claims based on the test scores" (Kane 2013, p. 1). Following this definition, validity is not the property of a test but a "property of tests for a particular kind of interpretation" (Kane 2013, p. 35). Therefore, validity arguments need to legitimize or delegitimize expected and unexpected interpretations of test scores (or PLM). Expected interpretations are to be understood as prior statements by future users of derived scores.

The term validity covers a variety of verification procedures that help locate a test regarding its underlying theory and its interpretations (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education 2014). The validator evaluates one or several validity aspects and can postulate that the interpretation of a test (e.g., the interpretation of its test scores) is valid regarding these aspects. The sum of all these evaluations can then lead to an overall validity judgment for this specific interpretation of the test. In this context, interpretations and purposes should be formulated in a validity argument before the actual validation takes place. Multiple attempts have been made by researchers to build a framework for validation (e.g., Cronbach and Meehl 1955; Kane 1994; Lissitz and Samuelsen 2007). We located our validation study in a framework specifically designed for the validation of interpretations and usages of outcomes from standard-setting studies (Kane 1994; Pant et al. 2013).

Standard-setting and validity

The term standard-setting covers "a variety of consensual approaches that consist of having committees of experts set discrete cut scores on continuous proficiency scales" (Pant et al. 2009, p. 96). During a standard-setting workshop the judgments of experts lead to cut scores that subdivide a continuous proficiency scale into PL. The resulting PLM are used to report the results of LSA to educators, education policy stakeholders, and the general public. As the PL are related to content descriptions, they can be understood as a translation of a continuous scale to policy decisions on, for example, which educational standards a student needs to fulfill at a certain point in the school career.

Regardless of its importance in the described communication process, the standard-setting process incorporates subjective judgments, psychometrics, policy goals, and social desirability (McGinty 2005; Pant et al. 2010; Tiffin-Richards et al. 2013) making it a highly controversial enterprise (Sireci et al. 2009). Since standard-setting is a consensual procedure among experts; the resulting cut score is the most appropriate cut score in these conditions. How appropriate the cut scores are regarding their usage and interpretations can be shown in validation studies (Kane 1994).

Validity in the context of interpretations and usages of standard-setting outcomes has been challenged in previous studies (Plake and Cizek 2012). Kane (1994) introduced comprehensive evaluation elements for passing scores based on performance standards that were summarized and categorized into an overview by Pant et al. (2009). Evaluation elements can be classified into three evidentiary and one consequential element. The evidentiary elements are procedural, internal, and external (see Table 1). As the setting of cut scores is "a critical gateway between evidentiary […] and consequential

Table 1 Evaluation elements and aspects of standard-setting procedures. Adapted from Pant et al. (2009)

Evaluation element	Description of our validation
Procedural	
Explicitness	Clear and explicit communication within the process
Implementation	Clear, systematic, and rigorous implementation of the procedures as well as the selection and training of the participants
Feedback	Experts' confidence in the process and the resulting cut scores
Internal	
Consistency within method	Variability of the estimates of the cut scores
External	
Comparison to other sources of information	Relationship between decisions based on cut scores and lessons taught in science, mathematics, and German
Consequential	
Reasonableness of cut scores	Relationship between cut scores and intended as well as pursued studies in science and non-science fields

aspects of validation" (Pant et al. 2009, p. 97), all aspects of their future use and interpretation should be validated. The evidentiary aspects are associated with the relation of the cut scores to the underlying standards and theories. The consequential aspects are associated with the appropriateness of the PL given their purpose and interpretations in the assessment system (Kane 2008). Table 1 provides a brief description of the aspects that we have adapted to and investigated in our validation study (for a comprehensive description, see Pant et al. 2009).

Procedural validation documents the accurate execution of the standard-setting process. Studies stressing this validity element incorporate evaluations from the experts on the procedure as well as information on the selection of the experts. The experts evaluate the clarity of the standard-setting method and its implementation, i.e., the extent to which they understand the standard-setting procedure and their confidence in the cut scores (e.g., Freunberger 2013). *Internal validation* focuses on the consistency between panelists' judgments across rounds, groups, and methods. Evidence on this element is assembled, for example, by means of analyses of variance (Freunberger 2013) or generalizability analyses (Tiffin-Richards et al. 2013) and targets the reliability of products of a standard setting-procedure. The *external validation* element is based on comparisons with other sources of information, such as other standard-setting procedures (Tiffin-Richards et al. 2013; Sireci et al. 2009) or performance measures, and often neglected in standard-setting studies (Pant et al. 2009). Performance measures used in the past include standardized test results within the same domain, high-school grades, and course taking patterns (Sireci et al. 2009). For our investigation we used number of science lessons per week, which are also assumed to be related to proficiency in the corresponding domain. For example, the number of science lessons per week is expected to be strongly related to proficiency and hence PL. However, number of lessons in other domains is expected to be less or even negatively related to proficiency in science. In our study, the domains mathematics and science show a content-related proximity. In contrast, number of lessons in a domain not related to science, such as in the first language, are assumed not to be related to science achievement. Relating the PL not only to measures hypothesized as connected to the PL placement but also to measures hypothesized not connected to the PL placement is seldom done. If the interpretation and use of a PL placement withstands external validity, the PL can be interpreted as science PL that can be used to identify strengths and weaknesses in science at the end of upper secondary level: One of two components of information desired by the ministry stakeholders.

Consequential validation explicitly refers to the reporting, the usage, and the interpretation of cut scores and the resulting PL. These uses and interpretations can either be intended or unintended by the standard-setters. Consequential validation ensures that cut scores are adequately reported to stakeholders and that they use and interpret them accordingly. In connection with our standard-setting procedure in science, the second desired interpretation of the cut scores by the stakeholders concerned the extent to which students were prepared for science studies. So, the cut scores (as well as scale values) need to be related to this criterion. Therefore, students more proficient in science are assumed to be more likely to enroll in science studies and to pursue a career in science after graduation from university. Thus, relating the PL placement of students to their study intention and to their short-term study decisions will provide first insights into the validity of this inference.

Even though the theoretical debate focuses on inclusion of external and consequential checks in validity studies (Lane and Stone 2002; Kane 2013), studies dealing with these two elements are rare (Pant et al. 2009). Ultimately, the interpretations and uses of cut scores, that is PL, can only be as valid as the interpretations and uses of the underlying test (Haertel 2002). Hence, before the interpretation and usage of standard-setting outcomes are evaluated, the interpretation and usage of the underlying test need to be validated.

The Angoff method and two of its variations

In our study, we applied the Yes/No method (Impara and Plake 1997) and the extended Angoff method (Cizek and Bunch 2007). The widely used Angoff method has been adapted and differentiated into a variety of procedures in order to suit numerous contexts (Plake and Cizek 2012). It is a test-centered standard-setting procedure, in which experts estimate the probability that a minimally competent student or group of students on a specified PL will solve each item of a test correctly. The percentages are then aggregated for each PL. The Yes/No method requires experts to rate whether a minimally competent person on each PL would be able to answer each test item correctly (Plake and Cizek 2012).

The extended Angoff method is a parallel approach for polytomously scored items (Plake and Cizek 2012). For this variation experts rate how many points a minimally competent person will be able to score on each PL. To our knowledge, both variations have not been simultaneously applied to the same test in order to arrive at common cut scores.

Just like any standard-setting method, the two chosen methods have several advantages and disadvantages. Especially the (original) Angoff method has been criticized in a report on the standard-setting procedures for student achievement of the National Assessment of Educational Progress (NAEP; Shephard et al. 1993). The major critique was that estimating the percentage of minimally competent persons answering an item correctly is “an unreasonable cognitive task because judges have no basis for such judgements” (Shephard et al. 1993, p. 72). This critical view of the Angoff method was refuted by several standard setting experts (Cizek and Bunch 2007). This criticism led to a considerable number of comparison studies between the Angoff method variations and other mostly test-centered standard setting methods as well as to the development of the Yes/No method (e.g.; Hsieh 2013; Skaggs and Hein 2011; Yudkowski et al. 2008). In these comparison studies the (Yes/No) Angoff method has either been favored or showed no substantial differences to the other methods. For example, Skaggs and Hein (2011) compared the Yes/No Angoff method to the single-passage bookmark which is a variation of the original Bookmark method (Cizek and Bunch 2007). They found few differences in the cut scores that derived from both methods as well as in the experts’ perception of the standard-setting process and slightly more variation between experts within the Yes/No method. But the judgement within the single-passage bookmark method was perceived as more difficult.

Additionally, the Yes/No method has been favored by judges over the original Angoff method in standard-settings, because a dichotomous judgement seems to be easier for them than estimating percentages, which is the task within the original Angoff method (for the original method, Angoff 1971; for recent studies on the Yes/No method e.g., Yudkowski et al. 2008; Hsieh 2013). However, this simplification for the judges leads to a loss of information and cut score judgments tend to move from the center to the extremes (Hsieh 2013). Imagine a test that only consists of items a minimally competent student at PL1 would answer with a probability of 10%, at PL3 with a probability of 55%, at PL4 with a probability of 85%. All items would receive a 0 at PL1, a 1 at PL4 and a 1 at PL3. Nevertheless, the described mental game does not reflect the complex reality of a standard-setting procedure. It is more likely that the items within one given PL will vary in difficulty and the proposed mental game is not a realistic expectation for any set of test items (Hsieh 2013). Therefore the Yes/No method as well as probability-based methods will produce suitable cut-scores.

Statements for validity arguments

The present study reports on a standard-setting procedure that was applied to a science test administered at the end of upper secondary school. The test was taken from the National Educational Panel Study (NEPS), a project that, inter alia, aims to monitor science competence development across the lifespan in Germany (Hahn et al. 2013). The ministry stakeholders intended to interpret the results of our study in two ways. Mainly, the results were meant to provide information on the students’ average abilities and the student distribution across different PL in science in the final year of upper secondary schooling (Grade 13). The aim was to monitor the educational system of one federal state in Germany. Moreover, as the upper secondary school leaving certificate allows students to enter university, the stakeholders assumed that the placement of students on the PL would provide information regarding students’ ability to pursue science-related and non-science-related university studies. As both interpretations and uses of the tests were induced on the policy level, the results were only reported to the ministry stakeholders on an aggregated level and not to the individual students or schools.

We stated our hypotheses as validity statements alongside the evaluation elements (see Table 1). First, we were interested in the clarity and explicitness of our procedure and in the credibility of the derived cut scores. The statements for the *procedural validity* were:

P_1: The selection of participating experts leads to a variety of experienced experts in the field of standard-setting and item development.
P_2: The standard-setting experts evaluate the process as clear and explicit.
P_3: After an initial training, the experts feel familiar with the standard-setting process.
P_4: The experts show confidence in the resulting cut scores.

Validity statement P_1 is met when the panel consists of researchers in science education, science teachers, and ministry personnel as well as experts with various academic degrees. Validity statements P_2, P_3, and P_4 are met when the experts' averaged responses to these statements significantly exceed the mean of a rating scale.

The statement for the *internal validity* is:

I: The cut scores of the Yes/No method as well as the extended Angoff method are reliable and reliability increases from round 1 to round 2.

This validity statement is met when the generalizability coefficients for both methods are above 0.60 and when the coefficient of round 2 exceeds that of round 1.

Second, and regarding external validity, we intended to show that the placement of students on a specific PL was mainly due to their learning opportunities in science and not in other domains. Therefore, we related the placement on PL to the number of school lessons taught in various domains. The statement for the *external validity* is:

E: Number of lessons in science is a better predictor for students' placement on science PL than number of lessons in other domains.

The statement on external validity is met when the number of lessons in science (a) succeeds in discriminating between all science PL, (b) still discriminates after controlling for lessons in other domains, and (c) discriminates between the PL better than the number of lessons in other domains.

Finally, regarding consequential validity, we aimed to show that the placement of students across PL is connected to future science-related university studies, a relation that needs to be confirmed according to the second interpretation of the stakeholders. The statement for the *consequential validity* is:

C: The students' PL placement is positively related to choosing science studies in the future.

We strongly emphasize this consequential element, because we assume that it might be the "weakest part of the interpretive argument" (Lane and Stone 2002, p. 23) and, at the same time, the most ambitious interpretation of the cut scores and resulting PL by stakeholders (Kane 2013).

Method

Study design and participants

The LISA 6-Study (Educatioal Outcomes of Students from Vocational and Academic Upper Secondary Schools-Study) was carried out in one federal state of Germany in order to provide information to stakeholders about the abilities of students in upper secondary level education in mathematics, science and English (first foreign language). In spring 2013, the 30-min NEPS-test on scientific abilities for university freshmen (Hahn et al. 2013) was administered to 3641 13th graders (52.3% male, $Ø_{\text{age}}=20$) in a simple one-form design. The test does not focus on knowledge or abilities needed for science university studies but rather on "areas that are generally agreed to be of lifelong significance" (Hahn et al. 2013, p. 115): Matter, biological systems, technological systems, development, and interactions (knowledge of science) as well as scientific enquiry and scientific reasoning within the contexts health, environment, and technology. The students were striving to obtain an upper secondary school leaving certificate (the German *Abitur*). The students were either enrolled in an academic upper secondary school (*allgemeinbildendes Gymnasium*; $n=1360$), or in a vocational upper secondary school (*berufliches Gymnasium*; $n=2281$; for a description of the two different school types, see Parker et al. 2013). Both secondary schools prepare students for university studies. The data from vocational upper secondary schools covered the entire student population in Grade 13 of the federal state. The data from the academic upper secondary schools (17 out of 99 schools) stem from the sixth measurement point of a longitudinal study that started in fifth grade in 2004/2005. As common in LSA, this first measurement point sample was a multistage stratified cluster sample. From measurement point five on, the whole grade level was drawn. The data were weighted according to population frequencies ($N_{\text{weighted}}=9621$) so that they are representative of the student population of the federal state. Three months after their graduation, 693 students (of 1396 who participated in the follow-up study) sent back a questionnaire pertaining to their university studies or vocational training. A total of 317 students had already started studying at university or college and had also reported their intended studies before graduation.

As neither PL descriptions nor cut scores for PL exist for the upper secondary level (A-level graduation phase, grades 11 to 13), we held a 2-day standard-setting workshop in autumn 2013. All PLM in Germany for subjects taught in schools, including science education, comprise four cut scores (e.g., English for intermediate school leaving certificate, Stanat et al. 2016; science and mathematics for intermediate school leaving certificate, Pant et al. 2013). In order to assure connectivity with PL at the end of the lower secondary level in Germany, we set four cut scores resulting in five PL.

The four cut scores and five PL set by a panel of nine experts (six ministry experts, three science education experts; seven female experts) were based on the NEPS scientific abilities test comprised of 30 items. Since a simulation study showed that ten experts is a desirable panel size (Wu and Tzou 2015), we invited ten experts. However, one expert did not show up to the standard-setting workshop. Seventeen items had a MC response format and the remaining 13 items were CMC items, for which the students could receive four points each. The students could reach a maximum of 69 points, including 52 points by answering the CMC items correctly. The test comprised content knowledge and scientific inquiry as well as the contexts health, technology, and environment. The test showed an expected a posteriori/plausible value (EAP/PV)-reliability of 0.86.

The 2-day standard-setting workshop started with an introduction to standard-setting in general, to the two Angoff variations, and to the tasks of the following 2 days. The experts were also briefly trained in the standard-setting procedure and discussed expected proficiencies of students on each PL. Before giving the judgements during the first round, for each PL the experts developed a description of the abilities of the students on that PL. This description was the basis for the following setting of the cut scores. The experts then received a booklet containing the 30 science items. The 17 MC items were followed by the 13 CMC items. For the MC items, we applied the Yes/No Angoff method (Impara and Plake 1997); for the CMC items, we applied the extended Angoff method (Cizek and Bunch 2007). The experts went through the booklet imagining a minimally competent student on each PL and individually set the four cut scores. Following the Educational Standards in Germany, the resulting five PL are labeled: below minimum standard, minimum standard, normative standard, normative standard plus, and optimal standard. After this first round, the experts received and discussed two types of feedback on their group performance: the variability between their individual raw cut scores and the four average cut scores resulting from the individual judgements. During the second round, the experts individually went through their initial judgments and adjusted them if necessary. Subsequently, the experts received feedback on the percentages of students on each PL based on the average of their judgments. During a third round, the experts identified consensual cut scores through discussion on the cut scores of the first two review rounds. At the end of the discussion all experts agreed on the final cut scores.

To arrive at the overall cut scores, we first calculated the percentage of correct items (MC items) and of correct scores (CMC items) separately for each expert. This procedure led to 18 individual cut scores (two for each expert), which we averaged across all experts within each Angoff variation. Finally, we added these two cut scores across the two Angoff variations. On the second day, the experts developed the PLD (see Appendix A). The PLD were developed on the basis of the items on each PL. The experts could base the PLD on at least three items per PL with the average number of items per cut score being 7.5 items.

The experts were asked to fill in questionnaires at four measurement points: before and after the process as well as after each round. The questionnaires covered the clarity and explicitness of the process. They also focused on the experts' familiarity with the process and confidence in the resulting cut scores. The questions which were taken from published standard-setting questionnaires and translated into German could be answered on a four-point Likert scale (1 = *strongly disagree* to 4 = *strongly agree*). During the standard setting process, the missing PLD were substituted by a characterization of students' proficiency for each PL as a basis for setting the cut scores.

Statistical analyses

To support the *procedural validity*, we report the means of answers on the questionnaires regarding clarity and explicitness of the process (P_2), familiarity with the process (P_3), and confidence in the resulting cut scores (P_4). In addition, the selection and the expertise of the participants are described (P_1). We applied a t-test to determine whether the answers were significantly above (or below) the scale mean (2.5) as well as a repeated measurement ANOVA for changes from round 1 to round 2.

In order to tackle the *internal validation* element, we identified variance components (VC) and calculated generalizability coefficients (Tiffin-Richards et al. 2013): a coefficient that corresponds to the reliability coefficient in classical test theory (CTT) using the generalizability theory approach (Brennan 2001; I_1). Coefficients between 0.40 and 0.59 are moderately reliable and coefficients below 0.40 indicate insufficient reliability (Landis and Koch 1977). The generalizability theory is a liberalization of the ANOVA and the CTT (Brennan 1992) and makes it possible to disentangle multiple sources of error—in our case regarding the judgement of experts. These multiple sources of error partition the single, undifferentiated random error term within CTT.

We ran our multivariate balanced G study in G string IV (see Fig. 1; Bloch and Norman 2011). In our G study for MC items, i × e, 17 items (i) form the population (object of measurement) and nine experts (e) the condition of measurement (universe; see Appendix B for the formula). The data point for each item is the cut score that is met when a student answers this item. In our G study for CMC items, i × e (c:e), 13 items (i) form the population (object of measurement), and nine experts (e) and four cut scores (c) that are nested within experts (c:e) the condition of measurement (universe). Both analyses were performed for each of the two rounds. The generalizability results are enhanced by descriptive findings. We estimated the reliability-like coefficient *index of dependability* ϕ (see Appendix B for the formula), which allows for absolute interpretations, in our case, cut scores set by other experts (Bloch and Norman 2011).

We also conducted four consecutive multinomial logistic regressions in M*plus* 7 (Muthén and Muthén 1998–2010) to target the *external validation* element (E_1). The criterion for the regressions was the placement on the PL for each of the 3641 students. The three external school measures number of lessons per week in science, in mathematics, and in German (mother tongue) were the predictor for their PL placement. All three measures were successively included into the prediction of students' classifications into

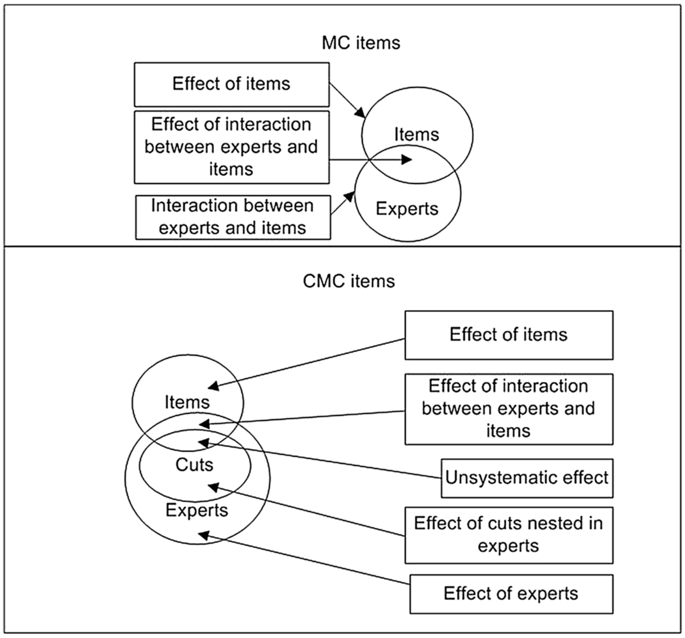
MC items
Effect of items
Effect of interaction between experts and items
Interaction between experts and items
Items
Experts
CMC items
Effect of items
Effect of interaction between experts and items
Unsystematic effect
Effect of cuts nested in experts
Effect of experts
Items
Cuts
Experts

Fig. 1 Decomposition into systematic and unsystematic error terms in our G studies for MC and CMC items. *MC* multiple choice

PL. Number of lessons varied substantially between the three domains. In mathematics and German, students either have three or five lessons per week in vocational upper secondary schools or four lessons per week in academic upper secondary school. Science is taught separately for biology, chemistry, and physics. Depending on the students' profile (e.g., science or languages), they have between two and ten lessons in up to three science subjects per week. We display the results for the difference between two consecutive cut scores, which corresponds to the log odd of this category versus the consecutive category. We also performed separate analyses with each category as the reference category to obtain the standard errors for these log odds.

We used their plausible values (PVs) to place students on PL regarding their science abilities. Please refer to the technical report of the PISA study (Organisation for Economic Co-operation and Development 2014) for more information on the general methodological procedure to calculate PVs. We used a partial credit model (Embretson and Reise 2000) that included an extensive background model (Leucht and Köller 2016).

In the first multinomial regression model, lessons per week in science predicted the classification of students into PL. Lessons in mathematics were included in the second model, and the third model included lessons in German. All lessons variables were *z*-standardized. We included school type as a predictor in Model C to account for differences between PL based on lessons dependent on school type.

We ran further regression analyses for the *consequential validation* that included study-related measures (C_1). We again used the student data on their intended and pursued studies as a proxy for the extent to which the students were prepared. Our analyses show whether students with strong abilities in science also opted for science studies, which would suggest that highly skilled students feel prepared. All students answered the open question: "Which field of study are you pursuing?" We coded the answers as a science (2), a science-related (1), or a (0) non-science study field. We retrieved the information regarding their study field 3 months after their graduation from a subsample of 317 students, who gave their consent to participate in a follow-up study. We coded this information accordingly.

We conducted missing value analyses between the students who did not report their studies after graduation (t1, $n = 3324$) and the students who did (t2, $n = 317$). We found no differences between the two groups regarding age, $t(381.242) = 0.880$, $p = 0.38$, and mother tongue (97% of the T2 students and 96% of the T1 students reported German as their mother tongue). However, we did find differences for proficiency in mathematics[1], $t_{\text{PV1}}(3639) = -4.431$, $p < 0.001$, and science, $\text{t}_{\text{PV1}}(3639) = -4.873$, $\text{p} < 0.001$, with students who reported their studies (t2) scoring higher in mathematics (25 points) and in science (29 points). T2 students also tended to have more books at home: an indicator of cultural assets at home. Smaller differences were found for sex ($\text{females}_{\text{t1}} = 54\%$, $\text{females}_{\text{t2}} = 61\%$) and school type ($\text{academic upper secondary}_{\text{t1}} = 37\%$, academic upper $\text{secondary}_{\text{t2}} = 40\%$). The comparison between the subsample and the overall sample shows that the subsample is quite small and slightly positively selected as well as less heterogeneous regarding the observed background variables. Therefore, this subsample

[1] The mathematics test contained dichotomous items only and was scaled using a Rasch model. The students could receive a maximum of 20 points on 20 items and the PLM incorporated the same number of cut scores.

Table 2 Cut scores of standard-setting procedure (average number of items solved at border of a cut score)

Item format	Average number of items round 1			Average number of items round 2		
	MC (*SD*)	CMC (*SD*)	All (*SD*)	MC (*SD*)	CMC (*SD*)	All (*SD*)
Cut 1	5.33 (2.55)	4.44 (2.08)	9.78 (4.20)	0.67 (1.19)	1.81 (0.91)	2.47 (1.80)
Cut 2	10.33 (2.38)	7.36 (1.95)	17.69 (3.60)	4.33 (2.55)	4.28 (1.56)	8.61 (3.00)
Cut 3	14.67 (2.72)	9.61 (1.82)	24.28 (3.60)	12.56 (1.70)	7.17 (1.56)	19.72 (2.40)
Cut 4	16.78 (0.51)	11.28 (1.69)	28.06 (1.80)	16.78 (0.51)	10.31 (1.95)	27.08 (2.40)

Number of items to be solved for MC and CMC items refers to number of items within the same item format

MC multiple choice, *CMC* complex multiple choice, *SD* standard deviation

is to some extent not representative and findings on the consequential validity should be interpreted with caution.

In separate regression analyses, we then regressed the placement on PL as well as the PVs in science and in mathematics, on the intended studies on the one hand, and on the pursued studies on the other hand. As mathematics showed content-related proximity, this measure was taken as an external criterion to evaluate the relative prediction of the science PL. German, was expected to predict placement on the science PL to a lesser extent.

Results

In our study, we investigated the validity of the stakeholders' two interpretations of the cut scores derived from our standard-setting procedure. We report the results in line with the validation framework of Pant et al. (2009). Table 2 shows the cut scores derived from the standard-setting procedure.

Obviously, the experts lowered their expectations regarding the three lowest cut scores within both variations. During the first round, the cut scores for MC items (derived from the Yes/No Angoff method) were set consistently higher than those for the CMC items (derived from the extended Angoff method). Since most standard deviations decreased, they show a general trend towards a consensus from round 1 to round 2 within and across both Angoff variations. Cut 4 is an exception. As almost all experts expected all students on PL4 to solve all of the items in round 1, the standard deviation increased once this expectation was lowered.

Procedural validity

The ministry experts were recommended by the Ministry of School and Professional Education following a formal request. The proposed experts had specific experience in curriculum development for upper secondary school and had already been involved in item development for LSA. The experts were one school supervisor (domain of expertise: electrical engineering), one expert from the state's institute for quality development of schools (physics), two directors of studies at schools (biology and chemistry), and one domain-specific supervisor (biology and metal technology).[2] The science education

[2] The information for one expert was missing because the expert did not fill out the questionnaire prior to the initial training.

Table 3 Experts' evaluation of the clarity and explicitness of the standard-setting process (Mean and standard deviation)

Statement	Overall	Round 1	Round 2
Introduction enabled clear understanding	3.78 (0.44)*	–	–
Training and exercises helped me	3.67 (0.87)*	–	–
Instructions were clear and explicit	–	3.56 (0.53)*	3.33 (0.50)*
I understood my task	–	4.00 (0.00)[a]	3.78 (0.44)*
Certainty about recommendations for final cut scores	2.89 (0.78)	–	–
I would sign a statement that recommends the usage of the cut scores	1.22 (0.44)*	–	–

Values can range between 1 = strongly disagree to 4 = strongly agree

– Not administered

* Significant at the level $p = 0.05$

[a] No variance, significance test not possible

experts worked in the fields of biology, chemistry, and physics education and were recruited from an institute with an excellent nationwide reputation in research on science didactics and empirical educational research that also operates locally in the schools (Leibniz Institute for Science and Mathematics Education). Two experts held assistant professorships and one was an associate professor. All science education professors had teaching experiences before pursuing an academic career; all five ministry experts, from which we received the information, worked at schools as part of their current post. To sum up, the expert recruitment resulted in a wide range of expertise regarding academic and vocational upper secondary schools (P_1).

The experts initially reported that they had understood the method and their tasks (P_3, see Table 3). In rounds 1 and 2 they also reported the explicitness and clarity of the process as significantly above the scale mean, which declined slightly yet not statistically significant throughout the process (P_2). Even though the experts demonstrated confidence regarding their duties, they did not show a comparably high degree of confidence regarding their output: the cut scores (P_4). On average, they would rather not have recommended the use of the cut scores and were not very certain about their final judgment.

Internal validity

The internal validity element relates to the reliability of the cut scores derived from the two different Angoff variations. In our four G studies (two rounds for each Angoff variation), we examined whether the error variance was explained by experts using differing criteria when setting standards (see Table 4).

The reliability for the Yes/No method dropped from $\Phi = 0.63$ in the first round to 0.48 in the second round. It remained the same for the extended Angoff method at $\Phi = 0.42$ (I_2). The rather low levels are moderate which showed that the experts did not use highly converging criteria when setting the cut scores (I_1). Table 4 shows that the variance component of the experts is highest in both methods and both rounds.

The generalizability theory explains the error within the experts' judgment. Our descriptive findings show whether the agreement between the experts increased,

Table 4 Science ability cut score VCs and G coefficients of the reliability analyses

Effect	*df*	MC items				Effect	*df*	CMC items			
		Round 1		Round 2				Round 1		Round 2	
		VC	%	VC	%			VC	%	VC	%
$\hat{\sigma}^2(\mathrm{i})$	16	0.179	47.4	0.063	48.1	$\hat{\sigma}^2(\mathrm{i})$	12	0.045	41.7	0.043	41.7
$\hat{\sigma}^2(\mathrm{e})$	8	0.105	27.8	0.069	1.5	$\hat{\sigma}^2(\mathrm{e})$	8	0.000	0.0	0.000	0.0
$\hat{\sigma}^2(\mathrm{ie})$	128	0.094	24.8	0.066	50.4	$\hat{\sigma}^2(\mathrm{ie})$	96	0.026	24.1	0.000	0.0
						$\hat{\sigma}^2(\mathrm{c:e})$	27	0.012	11.1	0.017	16.6
						$\hat{\sigma}^2(\mathrm{ic:e})$	324	0.025	23.1	0.043	41.7
$\hat{\sigma}^2(\mathrm{X_{ie}})$		0.378	100	0.198	100	$\hat{\sigma}^2\left(\mathrm{X_{ie(ce)}}\right)$		0.108	100	0.103	100
Φ		0.63		0.48				0.42		0.42	

$\hat{\sigma}^2(\mathrm{X_{ie}})$ and $\hat{\sigma}^2\left(\mathrm{X_{ie(ce)}}\right)$ are total scores of variance, $\hat{\sigma}^2(e)$=variance of experts, $\hat{\sigma}^2(i)$=variance of items, $\hat{\sigma}^2(ie)$=variance of the interaction between experts and items, $\hat{\sigma}^2(ce)$=variance of the cut scores that are nested in experts and $\hat{\sigma}^2(\mathrm{ic:e})$=variance of interaction between the items and the cut scores that are nested in experts

Φ index of dependability, *VC* variance components, *df* degrees of freedom

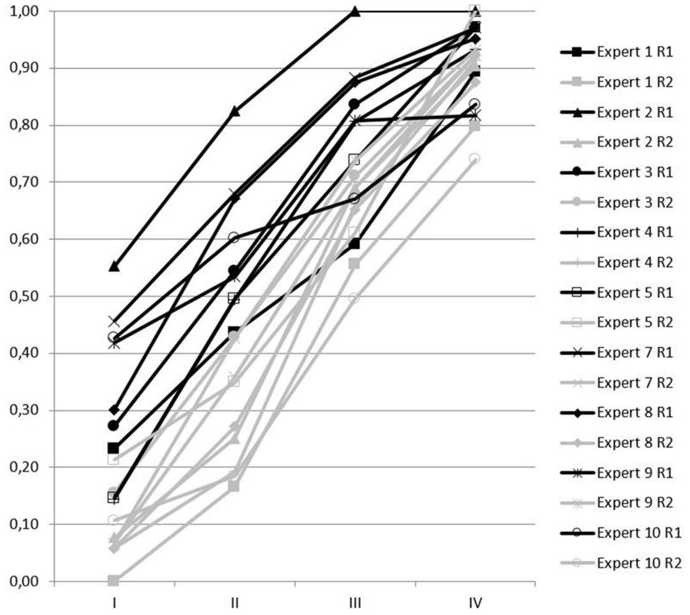

Fig. 2 Ratings of the experts for each cut score by round (integrated results). Values of the same round share the same color, values of the same expert share the same marker. Expert 6 did not show up at the workshop. *R* round

which means that the error decreased (see Fig. 2 as well as Appendix C for separate results of the two Angoff methods).

The variability of the cut scores declined from round 1 (*SD* between 4.20 and 1.80) to Round 2 (*SD* between 3.00 and 1.80). Hence, the experts' agreement did increase during the two rounds. However, the remaining error can still be explained by the same variance components.

Table 5 Multinomial regression of number of lessons per week on PL

Criterion	Predictors	Model A		Model B		Model C		Model D	
		OR	$SE_{\text{log odds}}$	OR	$SE_{\text{log odds}}$	OR	$SE_{\text{log odds}}$	OR	$SE_{\text{log odds}}$
0 vs. 1	Science	0.65	0.31	0.72	0.30	0.72	0.30	0.77	0.31
	Mathematics			0.83	0.11	0.84	0.11	0.93	0.07
	German					0.98	0.07	1.00	0.06
	School type							1.49	0.35
1 vs. 2	Science	0.58*	0.15	0.62*	0.15	0.63*	0.15	0.66*	0.15
	Mathematics			0.73*	0.05	0.73*	0.04	0.80*	0.04
	German					0.95	0.04	1.01	0.04
	School type							1.54*	0.16
2 vs. 3	Science	0.55*	0.08	0.55*	0.08	0.56*	0.09	0.58*	0.09
	Mathematics			0.85*	0.04	0.81*	0.04	0.84*	0.05
	German					0.91	0.06	1.05	0.06
	School type							1.65*	0.19
3 vs. 4	Science	0.53*	0.12	0.52*	0.12	0.53*	0.13	0.53*	0.14
	Mathematics			0.86	0.08	0.72	0.23	0.66	0.25
	German					0.81	0.28	1.00	0.39
	School type							2.00	0.75

For each cut score, the reference category of the dependent variable *PL in science* is the consecutive cut score. M*plus* does not provide Pseudo R^2 for multinomial logistic regressions. The Nagelkerke R^2 for the four models calculating manifest regressions in SPSS for all five PVs are: $R^2_{\text{Model A}} = 0.17–0.20$, $R^2_{\text{Model B}} = 0.21–0.23$, $R^2_{\text{Model C}} = 0.21–0.23$, $R^2_{\text{Model D}} = 0.22–0.24$. OR = odds ratio, *SE*log odds = standard error of the log odds

* $p < 0.05$

External validity

We consecutively ran four multinomial regression models by applying models incorporating the external criteria number of lessons per week in science (Model A), mathematics (Model B), German (Model C; see Table 5), and school (Model D). We only present the results for the difference between two consecutive cut scores and the associated standard errors (see "Method" section).

The four blocks show the four odds ratios (ORs) estimates: PL below 1 vs. PL1, PL1 vs. PL2, PL2 vs. PL3 as well as PL3 vs. PL4. An OR of 1 indicates that the predictor cannot discriminate between two categories, an OR above 1 indicates a higher chance of being placed in the reference category, and an OR below 1 indicates a higher probability of being placed in the indicated category. An OR of, for example, 0.55 (PL2 vs. PL3 in Model A) means that one more lesson in science decreases the chance of being classified into the reference category (PL2 rather than PL3) by the factor 0.55.

We found statistically significant ORs well below 1 for each category, except for PL0 vs. PL1[3], regarding number of lessons in science (Model A). This first model shows that science lessons were able to discriminate between the cut scores. In Model B we added the number of lessons in mathematics per week. On all PL, the ORs for mathematics were at least 0.11 higher than those for science. The science ORs increased but were still statistically significant. The ORs for mathematics were significant in two cases. The second model shows that science lessons discriminated between cut scores better than mathematics lessons, when controlled for mathematics lessons. In our third Model C,

[3] The ORs for PL0 vs. PL1 did not become significant for all predictors because only a few students were placed in the category PL0.

Table 6 Effects of PL placement and ability in science and mathematics on studies in science

Domain	Intended studies			Pursued studies		
	β	***SE***	**Pseudo *R*²**	***β***	***SE***	**Pseudo *R*²**
PL						
Science	0.18*	0.09	0.03	0.15	0.08	0.03
Mathematics	0.13	0.08	0.02	0.08	0.08	0.01
PV						
Science	0.19*	0.08	0.04	0.15	0.08	0.02
Mathematics	0.16	0.08	0.03	0.10	0.07	0.01

Results stem from individual regressions

PL proficiency level, *PV* plausible value, β standardized regression coefficient, *SE* standard error of the standardized regression coefficient, Pseudo R^2 from SPSS analyses = explained variance. All analyses were conducted with the 317 students who provided data

* $p < 0.05$

we added number of lessons in German. None of the ORs for lessons in German became statistically significant. The general pattern for lessons in science and mathematics did not change in this model. After controlling for lessons in mathematics and German, science lessons were still a good discriminator, whereas lessons in German were not (E_1). We added school type in Model D and the picture stayed roughly the same. The Nagelkerke R^2 of the four models increased slightly when number of lessons in mathematics was added and stayed mostly the same when lessons in German and school type were added. This again shows that lessons in science are a good predictor for placing students on the PL in science (E_1). While lessons in mathematics helped to further explain this placement, lessons in German did not.

Consequential validity—placement in PL and intended and pursued science-related studies

Our further analyses focused on the appropriateness of the stakeholders' assumption that a higher PL placement in science shows that students are prepared for science-related studies. In our first block of four regression analyses, we regressed the PL placements in science and mathematics on intended and pursued science-related studies (see upper block of Table 6). The PL placement in science predicted intended and pursued science-related studies not strongly, but statistically significant ($\beta_{\text{intended}} = 0.18$, $\beta_{\text{pursued}} = 0.15$), accounting for 3% of the variance in both intended and pursued studies. The related Pseudo R^2 was not significant. The prediction of the PL placements in mathematics showed even lower coefficients and less explanation of variance.

As we have already stated, the validity of interpretations of the PL relies on the validity of the interpretation of the test scores, which are the basis for the standard-setting procedure. We reran the same four analyses with the continuous PVs of the science and the mathematics test (see lower Block of Table 6) to make sure that the poor consequential validity is not due to the interpretation of the test itself.

The patterns of the coefficients for the PVs are similar; only the regression of the science PV on intended science-related studies was statistically significant, the corresponding R^2 was not. To sum up, the analyses show that the PL placement does not substantially predict intended and pursued science-related studies (C_1).

Discussion

In our study, we investigated validation arguments that derived from a standard-setting procedure based on two variations of the Angoff method (Angoff 1971). By including all elements of the framework for standard-setting validation (Pant et al. 2009), we were able to address both desired interpretations of the cut scores by ministry stakeholders. This detailed investigation included external criteria as well as future interpretations of PL. Based on separate examinations of the four evaluation elements procedural, internal, external, and consequential validity, we will draw conclusions about the overall validity of the interpretations of the science PL.

By incorporating all four validity elements, we go beyond studies on procedural and internal elements only. These studies do not show whether certain interpretations and usages of cut scores are appropriate. On the one hand, our study gives the reader insights into how to meet new challenges that arise with increasingly complex assessments and standard-settings. On the other hand, our study underlines the importance of focusing validation studies on interpretations and usages as well as on consequential aspects of assessments and standard-setting.

Before we elaborate on the evaluation elements, we need to point out some general limitations. Standard-setting in the context of educational assessment remains a complex largely unstandardized and critical element of testing (Bejar 2008). First, numerous suggestions exist about which method to choose and how to select materials and experts (e.g., Arrasmith and Hambleton 1988; Cohen et al. 1999). In our study, we chose the Angoff method, because it can easily be applied to different item formats of a test (Plake and Cizek 2012) and showed similar results compared to other (test-centered) standard-setting methods (e.g.; Hsieh 2013; Skaggs and Hein 2011; Yudkowski et al. 2008). Any choice of a standard-setting method leads to advantages and disadvantages, which we have laid out in the theoretical part of this article. When reviewing the literature, we observed that most comparison studies focus either on test-centered or person-centered methods. Because test-centered methods involving human judgement are particularly criticized, comparison studies should be expanded to incorporate both approaches. Our choice of standard-setting method in itself might already call for a validation study. Standard-setting procedures and our approach in general have several limitations. First, standard-setting processes involve social processes such as discussions which can only be captured with qualitative approaches. Since one expert did not agree to audiotaping the discussion we could not incorporate our qualitative data. Second, we applied four cut scores to a relatively short science abilities test. Third, our validation study does not incorporate the investigation of whether or not the level of the cut scores are valid in a normative sense.

In order to overcome these three main general limitations of standard-setting procedures, future studies should incorporate qualitative data on social processes during a standard-setting procedure and validity elements on the normative nature of PL. In our study, we used four cut scores in order to connect our PLM to existing PLM within the educational monitoring system in Germany (e.g., PLM for science education, Pant et al. 2013). All PLM have the PL below minimal standard, minimal standard, norm standard, norm standard+, and optimal standard.

The Angoff method as applied in our study has two method-specific limitations. First, within the framework of the original Angoff method (Yes/No method), experts have to give an in-our-out voting (1 or 0 for each task), a voting that calls for an extreme decision and might bias cut scores. Second, we applied two Angoff methods to the same test but to different tasks within the test, a procedure that was necessary given the nature of the items.

We nevertheless decided for the Angoff method, because experts consistently state that they find this method easier compared to others (e.g., Skaggs and Hein 2011; Yudkowski et al. 2008) and it bears the chance to apply different variations to different items of the same test, in our case the Yes/No and the extended Angoff method. Furthermore, we decided for four cut scores, because any other number of cut scores would not be integrable into the German educational monitoring system. In addition, the experts had at least three items (cut score 1, round 2) and a maximum of 11 items (cut score 3, round 2) to describe the resulting PL. The average number of items per cut score was 7.5 items.

Because this validity aspect was not the focus of our study, future validation studies should apply multiple standard-setting methods and compare the processes as well as the results in order to validate the choice for a specific standard-setting method (e.g., for assessment in English education Hsieh 2013; for university education Çetin and Gelbal 2013; for medical education Yousuf et al. 2015).

Procedural validation

The description of the expert panel for the standard-setting showed that the experienced experts represent the diversity of the field (P_1). The standard-setting experts evaluated the standard-setting process as clear and explicit (P_2) and felt familiar with the process (P_3). However, they did not show the corresponding confidence in the resulting cut scores (P_4). As this lack of confidence could impair the interpretation of the cut scores, the cut scores should be interpreted with caution. One reason for the experts' lack of confidence in the cut scores could have been the tight time schedule of 2 days. The limited time might not have left enough room for the experts to adjust their normative expectations to the lower abilities of the students. Moreover, PLD were unknown beforehand, which could have reduced the confidence when setting cut scores. Future standard-setting procedures should provide extra time for this process of adjustment between norms and reality and extra time in case the PLD are not known beforehand. Finally, even though former studies showed that experts found setting cut scores within the Yes/No method easier (e.g., Yudkowski et al. 2008), we cannot rule out that experts might have had more confidence in their tasks, when setting cut scores through another standard-setting method.

Very often, PLD already exist before the standard-setting procedure commences and experts can use them as a guideline for the cut score setting. This was not true in our case. The experts did develop temporary PLD during the introductory and training phase and determined a characterization of students for each PL. We assume that this procedure might have led to the remaining positive results (P1–P3). Therefore, this procedure should be applied to future standard-setting procedures that have the same prior conditions and should be looked into more specifically.

Internal validation

Our G studies showed that the experts made cut score recommendations across the two Angoff variations, of which the reliability was between 0.40 and 0.60 (I_1). Therefore, the future usage of the science PL for Grade 13 can only be called moderately internally valid. The variability of the cut scores within the expert group declined from round 1 to round 2 for the Yes/No method (MC items) and stayed the same for the extended Angoff method (CMC items). The combination of these two methods should be evaluated in light of the fact that the experts did not recommend usage of the cut scores. It again shows that more time for the process of agreement and convergence might have led to more reliable cut scores. Both results alter the validity and call for a cautious usage and interpretation of the cut scores. Further research on validation of the interpretation of cut scores deriving from the application of different standard-setting variations or methods should address the new challenges of these complex procedures.

External validation

To gain insights into the interpretation of the science PL, we calculated four consecutive multinomial logistic regressions that included numbers of lessons per week in science, mathematics, and German, as well as school type. Our results show that number of science lessons indeed discriminated between all science PL, even after controlling for the number of lessons in other domains (E_1). In the light of the test focusing on lifelong learning rather than the curriculum, significant OR that are quite substantial ($OR_{difffrom1} < 0.33$) show that the PL reflect proficiency in science and can be used accordingly. An exception was the cut score between PL below 1 and PL 1, which might be due to the low distribution of students on the former PL. The interpretation and use of all other PL were externally valid. Since reading achievement is a necessity for learning in other subjects and helpful to answer tasks that include written text, it might seem surprising that this predictor did not relate to the science achievement cut scores. Reading is an ability that students learn in school. But throughout the school career they will encounter various out-of-school situations (e.g., reading books) in which they apply their abilities and expand on them. This is not as extensively the case for science and mathematics. Therefore and in opposition to grade in German or a reading achievement assessment, number of lessons in German might not be related to the science achievement cut scores.

By using number of lessons in school subjects as an external validation criterion, we took one of several possible approaches to investigate external validity. Another prominent approach is the application of two different methods in order to compare results of both methods. Additionally, we only took into account a few of all the possible external criteria. A comprehensive external validation of the interpretations and usages of PL would be more than sufficient for one single validity study. Because addressing only one validation element is not in line with the comprehensive view of current validation theory (Messick 1994), we needed to rely on examples of external criteria. We are nevertheless convinced that number of lessons per week is a quite powerful criterion. Students receive most of their learning opportunities in the school environment, particularly in upper secondary level. However, the measure we applied did not incorporate instructional quality of lessons, the specific contents and skills that were taught or different teaching styles. These might differ for different school types or even between schools.

Our study took quite a broad approach to external validity. Future validation studies (of interpretations and usages of standard-setting outcomes) should also opt for such a broad perspective in order to define how much variance can be explained by close and distal external criteria. This approach provides strong support in favor of a content-related interpretation of cut scores.

Consequential validation

Evidence of the consequential validation provided valuable indications for how to interpret the PL and, maybe more importantly, for how not to interpret them. So far, validation studies incorporating consequential aspects have mainly focused on their media coverage (Hambleton and Meara 2000) and on stakeholders' perceptions of assessment results (Bullock and DeStefano 1998). We connected our validation study to interpretations that were formulated by stakeholders, such as: Are the students prepared for university science studies? By demonstrating that the interpretation connected to the consequential aspect is inappropriate, we prevent future misinterpretation of the PL. The cut scores do not seem to be valid regarding this specific interpretation of the PL placements by stakeholders. The gap between desired interpretation of the PL (extent to which students are prepared for science studies) and results on the consequential aspect was communicated to the stakeholders. We point out that the consequential validity aspect does not refer to an actual consequence for test takers, i.e., students. Rather, the consequential aspect supports (or does not support) the consequential interpretation of the stakeholders. Since capability to succeed in science studies and wish to study or actual study of science are not congruent, our results on consequential validity should be interpreted with caution regarding this specific inference.

We see two further reasons why our analyses could not meet our expectations regarding consequential validity. First, our results of pursued studies is based on a small slightly positively selected subsample with more female students that tend to have a higher cultural capital and higher abilities in science and mathematics. In this sense, our results on the consequential validity aspect can only serve as a first exploration of this aspect.

Second, we used a scientific ability test. This test focuses on abilities for lifelong learning and daily life (Hahn et al. 2013). Therefore, the test that was the basis for the standard-setting procedure was not tailored to requirements of science studies at university. We also emphasize that a test on science proficiencies required for science studies would be difficult to administer. On the one hand, students at the end of upper secondary level show a wide range of abilities within one domain. On the other hand, the domain-specific abilities and knowledge required for science studies are quite extensive, demanding, and very specific. We feel that administering such a test would be inappropriate, because most students in non-science profiles are very likely to fail such a test.

In conclusion, our results suggest that the hypothesis that students' decision to commence science or non-science studies is directly connected to their placement in a PL based on an abilities test at the end of the 13th grade is an overestimation and inappropriate. Interpreting our cut scores in this way does not appear to be valid. This finding is in line with previous research on predictors for STEM-career choices (science, technology, engineering, and mathematics). Students do not seem to opt for science-related studies solely on the basis of their achievement and knowledge, but also on the basis of parental support, self-concept, and interests in science and mathematics (e.g., Wang et al. 2017).

The exploratory nature of our results on the consequential validity aspect shows that researchers need to plan the investigation of this aspect with special care. Questions such as what the appropriate measures are, or how the sample should be acquired, need to be considered in order to meet the challenges of incorporating this element into a validity study. Regarding similar future assumptions of stakeholders, researchers need to communicate the research design and costs connected with the validation of these assumptions.

Overall evaluation of the validity

Taking a look at the comprehensive body of validity evidence for the interpretation of the cut scores that derived from our standard-setting, we conclude that our validity study was fruitful in several respects. We were able to show how the cut scores should *and* should not be interpreted. They can—with caution—be used as an indicator of 13th graders' strengths and weaknesses in science. They should not be used as an indicator for preparedness for science university studies. Since assessment formats are continually evolving and thereby leading to more complex designs, more research needs to be conducted on the application of new standard-setting methods to meet challenges arising from that development (in our case, the combination of two methods that leads to combined cut scores). The fact that the cut scores should not be used as an indicator for preparedness for university science studies could be regarded as controversial. However, we think this is the most enlightening part of our results. Our analyses highlight the importance of all four elements in a validity study. In contrast to previous validity studies that focused on internal and procedural aspects (Hurtz and Auerbach 2003; Massey 1997; McGinty 2005), our study shows the relative importance of the external and consequential elements. If we had not incorporated those two elements, we could not have shown that the PL can be related to science learning opportunities but not to decisions for science studies—two desired interpretations by the ministry stakeholders. Future studies should include these two aspects as well as the political and societal normative aspects of cut scores. We hope to have generated an impulse to future validity studies on the interpretation of standard-setting and assessment outcomes to include this very important validity element into their framework. Furthermore, we hope to have emphasized the importance of thoroughly investigating new standard-setting approaches as a consequence of ongoing assessment developments.

Abbreviations

ANOVA: analysis of variance; CMC: complex multiple choice; CTT: classical test theory; *df*: degrees of freedom; EAP/PV: expected-a posteriori/plausible value; LSA: large scale assessments; MC: multiple choice; N: sample size; n: subsample size; NEPS: National Educational Panel Study; OR: odds ratio; PL: proficiency level; PLD: proficiency level description; PLM: proficiency level model; PV: plausible value; R: round; *SE*: standard error; VC: variance components.

Authors' contributions

NK carried out the standard setting procedure as well a the validation study. She wrote the different drafts including the final one. HW performed the G studies and provided feedback on the different drafts of the manuscript. OK provided feedback on the different drafts of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Table 7 Proficiency level descriptions

PL4 Optimal standard	Transfer basic concepts (principles, theories and laws) to other domains Flexibly apply concepts in various contexts Logically connect several variables
PL3 Norm standard plus	Take complex information and relations from a diagram and interpret the information Apply content knowledge and apply it in new contexts Apply basic concepts (principles, theories and laws) Take scientific information from a challenging, specialized text
PL2 Norm standard	Take information of diagrams in a target-oriented manner without interpretation and compare if necessary Reproduce simple content knowledge and utilize it in context Take scientific information from a simple, specialized text
PL1 Minimum standard	Extract scientific information presented in a simply formulated specialized text Objectively extract information from a diagram without interpretation Take scientific information from a text or picture when these provide redundant/equal information Verify the veracity of simple scientific statements based on a graphical representation
Below PL1 Below minimum standard	Fail the requirements of lower secondary education

Availability of data and materials
The data of the student LSA has not yet been made available to the public. The project members are currently preparing the data for secondary analyses and plan to publish them at the Research Data Centre (FDZ) at the Institute for Quality Improvement (IQB).

Funding
Not applicable.

Appendix A

See Table 7.

Appendix B

The linear model for the G study (i × e) for the MC items is

$\hat{\sigma}^2(X_{ie}) = \hat{\sigma}^2(i) + \hat{\sigma}^2(e) + \hat{\sigma}^2(ie)$ where $\hat{\sigma}^2(X_{ie})$ represents the total scores of variance and $\hat{\sigma}^2(ie)$ the residual error term, that is, the interaction of items with experts in the model.

The formula of the Φ-coefficient resulted in the following specifications:

$$\Phi = \frac{\hat{\sigma}^2(i)}{\hat{\sigma}^2(i) + \hat{\sigma}^2(e) + \hat{\sigma}^2(ie)}$$

where $\hat{\sigma}^2(ie)$ represents the total error term.

The linear model for the G study, i × e (c:e), for the CMC items is

$\hat{\sigma}^2\left(X_{ie(c:e)}\right) = \hat{\sigma}^2(i) + \hat{\sigma}^2(e) + \hat{\sigma}^2(c:e) + \hat{\sigma}^2(ie) + \hat{\sigma}^2(ic:e)$ where $\hat{\sigma}^2\left(X_{ie(c:e)}\right)$ represents the total scores of variance and $\hat{\sigma}^2(ic:e)$ the residual error term, that is, the interaction of all facets in the model.

The formula of the Φ-coefficient resulted in the following specifications:

$$\Phi = \frac{\hat{\sigma}^2(i)}{\hat{\sigma}^2(i) + \hat{\sigma}^2(e) + \hat{\sigma}^2(c:e) + \hat{\sigma}^2(ie) + \hat{\sigma}^2(ic:e)}$$

where $\hat{\sigma}^2(ic:e)$ represents the total error term.

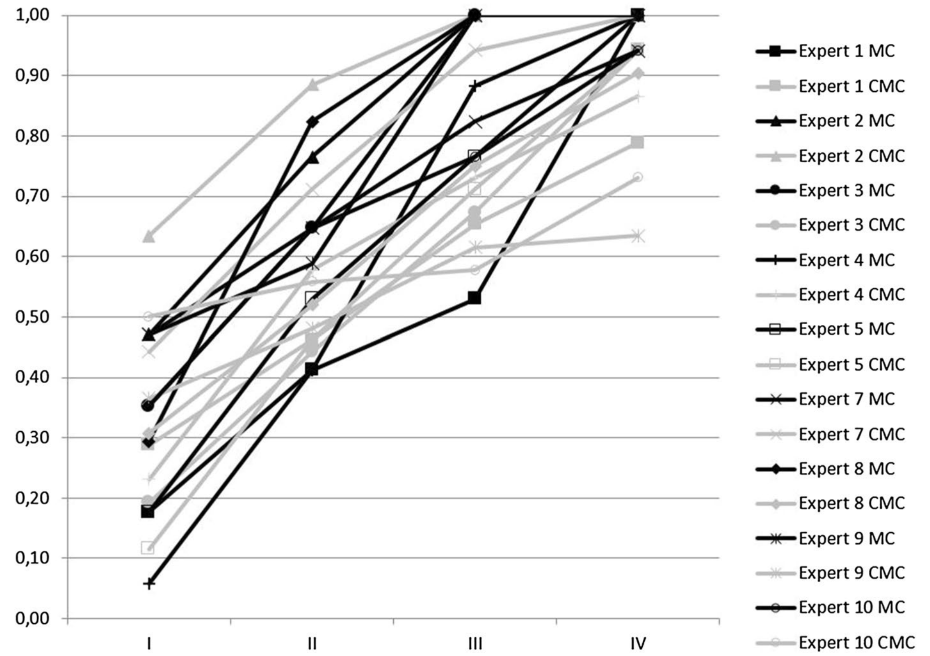

Fig. 3 Ratings of the experts for each cut score by Angoff method, round 1. Values of the same item format share the same color, values of the same expert share the same marker. Expert 6 did not show up at the workshop. *MC* multiple choice items, *CMC* complex multiple choice items

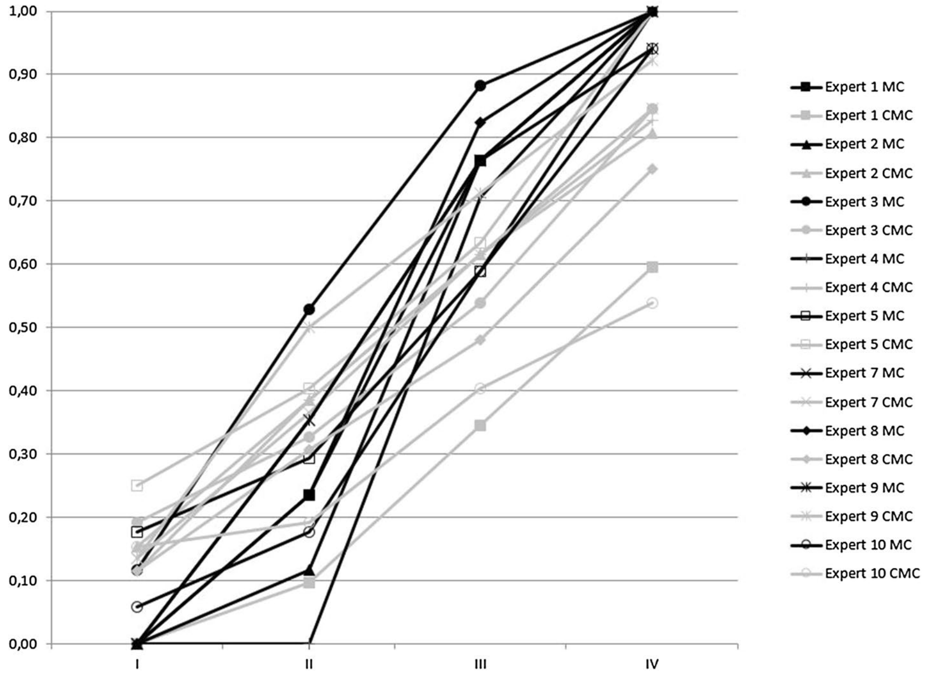

Fig. 4 Ratings of the experts for each cut score by Angoff method, round 2. Values of the same item format share the same color, values of the same expert share the same marker. Expert 6 did not show up at the workshop. *MC* multiple choice items, *CMC* scomplex multiple choice items

Appendix C

See Figs. 3, 4.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 8 June 2018 Accepted: 12 February 2019
Published online: 18 February 2019

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Arrasmith, D. G., & Hambleton, R. K. (1988). Steps for setting standards with the Angoff method. Retrieved December 01, 2017, from http://files.eric.ed.gov/fulltext/ED299326.pdf.

Bejar, I. I. (2008). Standard setting: What is it? Why is it important? *R & D Connections, 7,* 1–6.

Bloch, R., & Norman, G. (2011). G String IV User Manual. Ralph Bloch & Geoff Norman: Hamilton, ON.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27–33. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer Verlag.

Bullock, C. D., & DeStefano, L. (1998). A study of the utility of results from the 1992 results from the 1992 Trial State Assessment (TSA) in reading for state-level administrators of assessment. *Educational Evaluation and Policy Analysis, 20,* 47–51.

Çetin, S., & Gelbal, S. (2013). A comparison of bookmark and Angoff standard setting methods. *Educational Sciences: Theory & Practice, 13,* 2169–2175. https://doi.org/10.12738/estp.2013.4.1829.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting. A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage Publications.

Cohen, A. S., Crooks, T. J., & Kane, M. T. (1999). Designing and evaluating standard-setting procedures for licensure and certification tests. *Advances in Health Sciences Education, 4,* 195–207. https://doi.org/10.1023/a:1009849528247.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–301.

Ebel, R. L. (1972). *Essentials of educational measurement*. Engelwood Cliffs: Prentice Hall.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum Associates.

Freunberger, R. (2013). Standard-Setting Mathematik 8.Schulstufe. Technischer Bericht [Standard setting in mathematics Grade 8. Technical Report]. Retrieved December 01, 2017, from https://www.bifie.at/system/files/dl/StaSett_M8_TechReport_sV__2013-05-15.pdf.

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational measurement: Issues and practice, 21*(1), 16–22. https://doi.org/10.1111/j.1745-3992.2002.tb00081.x.

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., et al. (2013). Assessing scientific literacy over the lifespan: A description of the NEPS science framework and the test development. *Journal of Educational Research Online, 5*(2), 110–138.

Hambleton, R. K., & Meara, K. (2000). Newspaper coverage of NAEP results 1990 to 1998. In M. L. Bourque & S. Byrd (Eds.), *Student performance standards and the National Assessment of Educational Progress: Affirmations and improvements* (pp. 131–155). Washington, DC: National Assessment Governing Board.

Hsieh, Mingchuan. (2013). Comparing Yes/No Angoff and bookmark standard setting methods in the context of English assessment. *Language Assessment Quarterly, 10,* 331–350. https://doi.org/10.1080/15434303.2013.769550.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgement consensus. *Educational and Psychological Measurement, 63,* 584–601. https://doi.org/10.1177/0013164403251284.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34,* 353–366. https://doi.org/10.1111/j.1745-3984.1997.tb00523.x.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425–461. https://doi.org/10.3102/00346543064003425.

Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher, 37,* 76–82.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73. https://doi.org/10.1111/jedm.12000.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational measurement: Issues and practice, 21*(1), 23–30. https://doi.org/10.1111/j.1745-3992.2002.tb00082.x.

Leucht, M., & Köller, O. (2016). Anlage und Durchführung der Studie [Research design of the study]. In M. Leucht, N. Kampa, & O. Köller (Eds.), *Fachleistungen beim Abitur: Vergleich allgemeinbildender und beruflicher Gymnasien in Schleswig-Holstein* (pp. 79–98). Münster: Waxmann.

Leucht, M., Kampa, N., & Köller, O. (2016). *Fachleistungen beim Abitur: Vergleich allgemeinbildender und beruflicher Gymnasien in Schleswig-Holstein [Abilities at the end of upper secondary education. A comparison between academic and vocational upper secondary schools in Schleswig-Holstein]*. Münster: Waxmann.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36,* 437–448.

Massey, A. J. (1997). Multitrait-multimethod/multiform evidence for the validity of reporting units in national assessments in science at age 14 in England and Wales. *Educational and Psychological Measurement, 57,* 108–117. https://doi.org/10.1177/0013164497057001007.

McGinty, D. (2005). Illuminating the "black box" of standard setting: An exploratory qualitative study. *Applied Measurement in Education, 18,* 269–287. https://doi.org/10.1207/s15324818ame1803_5.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23. https://doi.org/10.3102/0013189x023002013.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749. https://doi.org/10.1037/0003-066x.50.9.741.

Muthén, L. K., & Muthén, B. O. (1998–2010). Mplus user's guide (7th ed.). Los Angeles, CA: Muthén & Muthén.

Organisation for Economic Co-operation and Development. (2014). *PISA 2012: Technical report*. Paris: OECD Publications.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation, 35,* 95–101. https://doi.org/10.1016/j.stueduc.2009.10.008.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.). (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I [IQB state comparison 2012. Competencies in mathematics and science at the end of secondary level I]*. Münster: Waxmann.

Pant, H. A., Tiffin-Richards, S. P., & Köller, O. (2010). Standard-Setting für Kompetenztests im Large-Scale-Assessment [Standard setting for competence tests in large-scale assessments]. *Zeitschrift für Pädagogik, Beiheft, 56,* 174–187.

Parker, P. D., Marsh, H. W., Lüdtke, O., & Trautwein, U. (2013). Differential school contextual effects for math and English: Integrating the big-fish-little-pond effect and the internal/external frame of reference. *Learning & Instruction, 23,* 78–89. https://doi.org/10.1016/j.learninstruc.2012.07.001.

Plake, B. S., & Cizek, G. J. (2012). The modified Angoff, extended Angoff, and Yes/No standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards. Foundations, methods, and innovations* (pp. 181–253). New York: Routledge.

Shephard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Stanford: National Academy of Education.

Sireci, S. G., Hauger, J. B., Wells, C. S., Shea, C., & Zenisky, A. L. (2009). Evaluation of the standard setting on the 2005 grade 12 national assessment of educational progress mathematics test. *Applied Measurement in Education, 22,* 339–358. https://doi.org/10.1080/08957340903221659.

Skaggs, G., & Hein, S. F. (2011). Reducing the cognitive complexity associated with standard setting: A comparison of the single-passage bookmark and Yes/No methods. *Educational and Psychological Measurement, 71,* 571–592. https://doi.org/10.1177/0013164410386948.

Stanat, P., Böhme, K., Schipolowski, S., & Haag, N. (2016). IQB trends in student achievement 2015. The second national assessment of language proficiency at the end of the ninth grade. Summary. Münster: Waxmann. Retrieved November 19, 2018, from https://www.iqb.hu-berlin.de/bt/BT2015/Bericht.

Tiffin-Richards, S. P., Pant, H. A., & Köller, O. (2013). Setting standards for English foreign language assessment: Methodology, validation, and a degree of arbitrariness. *Educational Measurement: Issues and Practice., 32*(2), 15–25. https://doi.org/10.1111/emip.12008.

Wang, M.-T., Ye, F., & Degol, L. J. (2017). Who chooses STEM careers? Using a relative cognitive strength and interest model to predict careers in science, technology, engineering, and mathematics. *Journal of Youth and Adolescence, 46,* 1805–1820. https://doi.org/10.1007/s10964-016-0618-8.

Wu, Y.-F., & Tzou, H. (2015). A multivariate generalizability theory approach to standard setting. *Applied Psychological Measurement, 39,* 507–524. https://doi.org/10.1177/0146621615577972.

Yousuf, N., Violato, C., & Zuberi, R. W. (2015). Standard setting methods for pass/fail decisions on high-stakes objective structured clinical examinations: A validity study. *Teaching and Learning in Medicine, 27,* 280–291. https://doi.org/10.1080/10401334.2015.1044749.

Yudkowski, R., Downing, S. M., & Wirth, S. (2008). Simpler standards for local performance examinations: The Yes/No Angoff and whole-test Ebel. *Teaching and Learning in Medicine, 20,* 212–217. https://doi.org/10.1080/10401330802199450.