Gauging the Instructional Sensitivity of Ordinal Assessment Items

Özge Altıntaş, Faculty of Educational Sciences, Ankara University

Yu-Hui Chang, Institute of Education, National Sun Yat-sen University

Setlhomo Koloi-Keaikitse, Department of Educational Foundations, University of Botswana

Anne Traynor, College of Education, Purdue University

* All co-authors contributed equally to our report

March 2025

Abstract

For achievement test scores to provide useful information about learning progress, differences in the implemented curriculum should affect student test-takers' item responses. Test item scores are "instructionally sensitive" if they are affected by variation in the content or quality of instruction received by students. This study was designed to inform item-writing guidelines for educational tests that assess students' current achievement or learning progress on school curriculum objectives. We used sequential explanatory mixed methods to characterize patterns in the content features of instructionally (in)sensitive constructed-response science test items and their scoring rubrics. In the first phase, our team applied generalized partial credit models to existing test item response data to identify items that have high and low instructional sensitivity. In the second phase, we guided master science teachers from four national education systems—Botswana, Chinese Taipei, Türkiye, and the United States—through item content analysis interviews to evaluate those items. By relying on the expertise of educators who are knowledgeable about both the item content and cognition of students at the tested grade level, we have gained new insights about the characteristics of instructionally (in)sensitive items, which we summarize at the end of this report.

1. Background Information

Achievement test item scores are “instructionally sensitive” if they are affected by variation in the content or quality of instruction received by test-takers (Muthén et al., 1991). When test scores will be interpreted as measuring learning by students or trainees in an education context, evidence that the test-takers have had adequate opportunity to learn the tested content is a crucial type of validity evidence (AERA, APA, & NCME, 2014). Without such evidence, the scores can be argued to measure test-takers’ current achievement status or predict their future outcomes, but achievement score differences across individuals or groups cannot be directly attributed to their learning in school.

Instructional sensitivity indices are statistics that measure how much students' test scores change when the instructional curriculum changes (Polikoff, 2010). Test item instructional sensitivity indices connect the educational policy concept of opportunity to learn (e.g., Moss et al., 2008) to empirical differences in test item difficulty (Naumann et al., 2020). An item with instructionally-sensitive scores has item difficulty statistics that vary for test-taker groups who have documented differences in opportunity to learn about the content topic tested by that item, but are otherwise equal. Although we define instructional sensitivity as a property of item scores in a test-taker population, rather than of items solely, to emphasize that our definition pertains to items, not only total test scores, we will use the term *item instructional sensitivity* (IS) throughout this report.

1.1. Tentative Principles for Writing Instructionally-Sensitive Achievement Test Items

Previous research allows us to extract tentative principles for writing instructionally sensitive multiple-choice test items. Items that require disciplinary knowledge to reach a correct solution tend to show relatively high IS (regarding mathematics items: Li et al., 2017; Muthén et al., 1991; regarding science items: Naumann et al., 2019). Conversely, items that can be answered correctly using “everyday knowledge,” which is often learned outside of school, seem to have lower IS (Li et al.). Expectedly, items framed in contexts that were not used during instruction also tend to exhibit lower IS (Ruiz-Primo et al., 2012). An earlier conclusion that seems troubling, if a test is intended to assess “deeper learning,” is that items requiring evaluation or synthesis appear to have low IS (Muthén et al.). For constructed-response items specifically, the content of the performance level descriptions in an item’s scoring rubric seems to be an important driver of the scores’ instructional sensitivity (Traynor, Li, & Zhou, 2025). However, the extent to which these item writing principles generalize to constructed-response items with scoring rubrics (e.g., Leighton, 2019) is largely uncertain.

Research seems to converge on the instructional sensitivity principles outlined above, but we judge their use for item writing as “tentative,” even for multiple-choice items, because it is not clear whether IS magnitude is a reproducible property of a particular item in a given test-taker population. Each existing IS study has analyzed item response data from only one assessment occasion across instructional contexts, or two time points in one instructional context. From our perspective, conclusions about item instructional sensitivity could be used to guide future item writing and test form assembly (e.g., Albano & Rodriguez, 2018; Zwick et al., 2018) if there is evidence that they are stable to a reasonable degree for a particular item (e.g., over time, across similar populations), but not otherwise.

1.2. Instructional Sensitivity Indices for Constructed-Response Test Items

Constructed-response test items that are scored in ordered categories include items assessed using a rubric (i.e., marking key) by content experts or an automated scoring algorithm. Each item's scoring system should consist of performance level descriptors that are aligned with educational curriculum objectives, directly observable in item responses, detailed enough to be consistently understood by raters, and separable (e.g., Brookhart, 2018). For constructed-response items, we conceptualized instructional sensitivity as a property of the scores yielded by a certain test item when test-takers' responses are rated using a particular rubric.

In recent research, Traynor et al. (2025) proposed and examined performance of an instructional sensitivity index for test items from a multiple-group generalized partial credit model (Muraki, 1999). They proposed that the difference between generalized partial credit item difficulty or step difficulty parameters in test-taker groups that have (1) and have not (0) experienced relevant topical instruction, "BDIFF" (Equations 1 and 2), might be a useful instructional sensitivity index. They concluded from the simulation results that the IS index was useful to evaluate item instructional sensitivity when at least 500 student test-takers who had, and who had not had, instructional exposure to the item's content topic took each test item. Criteria from Roussos et al. (1999) to evaluate the magnitude of the index value appeared to work well. As an alternative to this multivariate analysis method when student group sample sizes are small, an item IS index could be computed for one item at a time using multinomial logistic regression (Li et al., 2017).

1.3. How Much Item Instructional Sensitivity is Desirable? . . . And How Should We Measure Students' Opportunity to Learn in an Item IS Index?

Although achievement test development standards clearly recognize that students' opportunity to learn the tested content (or lack thereof) affects the validity and fairness of test score uses, the literature is more ambiguous about the desirability of item IS (Burstein, 1989; Darling-Hammond et al., 2013; Rodriguez, 2017). Its desirability depends on the conceptual "distance" between the student opportunity to learn variable and the test item score that are used to construct an IS index. We argue that test items should require students to demonstrate transfer of their learning to contexts that differ meaningfully from any examples presented during instruction. Content exposure effects on specific item performance would be undesirable if they reflect "teaching to the test" that narrows the school curriculum to emphasize incidental features of the test items (Koretz, 2017). Thus, measurement of students' opportunity to learn for an IS index should query broad content topic instruction, rather than instruction about exact item content or format.

Capturing variation in instructional content, for instance, as topic instructional time or topic exposure, is essential to measuring opportunity to learn that predicts student achievement test scores in experimental and observational studies (Floden, 2002; Schmidt & Maier, 2012). Conversely, variation in content-neutral instructional quality measures alone does not seem to affect classroom achievement test score mean differences (Ing, 2018). Thus, in this study, we will use content topic exposure indicators as a proxy for students' opportunity to learn. Another established result is that item instructional sensitivity indices tend to be lower for items that require transfer than for those that substantially reproduce the instructional content ("close" and "proximal 1" items; Ruiz-Primo et al., 2012, p. 701; see also Ruiz-Primo et al., 2002). Because transfer of learning to new contexts is a major goal of education, we argue that detectable IS of items that require transfer is desirable, so in this study, we will utilize response data from science test items that are curriculum-based, but not closely aligned with instruction.

1.4. Study Purpose

To compile instructional sensitivity evidence for constructed-response items, the first phase of this study applied an IS effect size index to multiple samples of Trends in International Mathematics and Science Study (TIMSS) science test item response data from eighth-grade students. Our purpose was to establish the degree of stability of item instructional sensitivity index values across different testing years for one test language version (similar student populations) and across language versions (different student populations). We asked: Do the results imply generalizable conclusions about the IS of any particular item? That is, does IS tend to be a stable property of test item scores? [Research Question 1].

After identifying constructed-response items and rubrics that produced scores with either high or low instructional sensitivity, we examined shared content features within the two sets of item/rubric pairs. We asked: What content features distinguish item/rubric pairs that yield item scores with negligible, and high, instructional sensitivity? [Research Question 2] During this second study phase, we conducted item content analysis interviews with master science teachers from four national education systems to explain patterns in the item IS index effect sizes. Then, we discerned themes and sub-themes from the item interview transcripts using reflexive thematic analysis. This research project was intended to inform test item writing and rubric design guidelines about instructional sensitivity, and allow improved control of the instructional sensitivity of test forms during their assembly.

2. Methods: Research Question 1

2.1. Data Source: TIMSS 2011 to 2019 Constructed-Response Test Items

Previous item IS studies have analyzed data from the United States. Because we hope to inform general test item writing principles, we sought a broader sample of item response data from eighth grade students who took the Trends in International Mathematics and Science Study 2011 or 2015 science assessment (TIMSS; Martin & Mullis, 2012; Martin et al., 2016) in English, the most administered language version of the test. In 17 countries (see Table 1), some or all test-takers took an English language version. We intentionally selected data from constructed-response items that had been administered during two different years and then retired from operational use, because interpreting item IS results requires examining the item content.

Students' item response data from international large-scale assessments like TIMSS is particularly useful for identifying items with high and low instructional sensitivity. Countries' intended science curricula are known to vary, which produces structural differences in students' opportunity to learn certain science content topics. These OTL differences are measurable, whereas OTL differences within one country or a classroom may be time-varying or more subtle. As explained later, calculating test item instructional sensitivity requires a clear operational (not only theoretical) definition of students' OTL. Further, high variation in individual students' OTL, like that observed across countries, allows degrees of item instructional sensitivity to be distinguished. The observational nature of international large-scale assessments also assures that the results reflect regular science instruction that students experience, rather than an atypical intervention curriculum.

The TIMSS study team collects instructional information at the individual student level, so the data is suitable for secondary analysis of the relationship between students' opportunity to learn and their achievement test item scores. Science teachers completed a questionnaire indicating, for each assessed topic (e.g., classification of living organisms, properties of matter), whether students had received instruction during a previous or current school year, or never. Up to three teachers of each student completed the questionnaire, so students in the same classroom could have distinct instructional histories reported. A student was classified as reported to have experienced topic instruction if at least one teacher indicated they had opportunity. As shown in Table 1, the proportion of individual students reported by their teachers to have received instruction ranged between .23 and 1.00 across topics and countries.

TIMSS implemented a two-stage stratified sampling design to enlist student participants, and their teachers (Martin et al., 2016). In total during each year, about 50,000 eighth graders (56,872 in 2011; 45,367 in 2015) from hundreds of schools (1,010 in 2011; 385 in 2015) across participating countries took an English language version of the science test. The items covered four content areas: Biology, Chemistry, Earth Science, and Physics. In our analysis, we used sampling weights that treated all countries as equal in size (i.e., "Senate" weights) so the results would not be driven by instruction-item performance relationships in the United States, which has a much larger youth population than the other 16 countries.

To compose our item set for the IS evaluation, we selected the 7 constructed-response items that were released in 2019, and then a random sample of 25 additional items, stratified by content area, from among those that were released in 2015 and also administered in 2011. Interpreting IS results requires examining the test item content, so 32 items was the maximum number that we could analyze, given the project resources. The 32 items were dispersed across multiple test booklets. From among the remaining items in each booklet, for each year, we again drew a stratified random sample so eight items per booklet, two from each content area, would be included in the student achievement score estimation.

Descriptive statistics for the student achievement test item and teacher instructional questionnaire data by country and overall are summarized in Table 1. Because the number of students in one of the opportunity-to-learn (OTL) groups who answered an item was often less than 30 within a country, we planned to analyze all responses to the English-version science test as one sample, and not to undertake within-country analyses. This analytic decision was also intended to maximize variation in students' OTL for each test item. At least 4,500 students produced written responses to each science item, as shown in Tables 2 and 3. Between 58% and 96% of students had reported opportunity to learn about each item's content topic. Fit statistics for a generalized partial credit model applied to the 8-item sets that we drew from each booklet suggested the data could be modeled as unidimensional (Tucker-Lewis index = [.93, .99]; Root mean square error of approximation = [.02, .07]).

2.2. Instructional Sensitivity Index

Item instructional sensitivity analysis can be conceptualized as a special case of uniform differential item functioning (DIF) analysis that compares groups with varying opportunity to learn the tested content. During the analysis, item difficulty statistics may be compared for test-taker groups before and after an instructional intervention, or groups whose instructional history is known to differ (Naumann et al., 2016). One important distinction between item IS and differential item functioning in general is that items exhibiting differential functioning are viewed as undesirable, potentially indicating problems with construct representation in the test scores (e.g., Gómez-Benito et al., 2018), while items with IS favoring the group that had more learning opportunity are desirable if the test scores are claimed to measure students' learning during school (Naumann et al., 2016; Ruiz-Primo et al., 2012).

Among differential item functioning methods for polytomous items, Muraki (1999) proposed testing group differences in an item's difficulty parameter, estimated from a multiple-group generalized partial credit model. It is conceptually well-suited as an item instructional sensitivity detection method because it can produce a group item difficulty difference index, which matches common definitions of item instructional sensitivity (Naumann et al., 2016; Polikoff, 2019). Also, the generalized partial credit model is often used for achievement test scoring. The multiple-group model parameterization that we have implemented is represented in Equation 1:

$$P(Y_{j,g} = k \mid \theta) = \frac{e^{\sum_{i=1}^{k} a_j(\theta - b_{j,g} + d_{i,j,g})}}{\sum_{h=0}^{m-1} e^{\sum_{i=1}^{h} a_j(\theta - b_{j,g} + d_{i,j,g})}}$$

$$d_{0,j} = 0$$

$$\sum_{i=1}^{m_j - 1} d_{i,j,g} = 0 \qquad (1)$$

where $P(Y_{j,g} = k|\theta)$ = Conditional probability of a response ($Y_{j,g}$) on Item j in Category k (k =0, 1,…, m – 1; m = number of observable values) for test-takers in opportunity-to-learn (OTL) Group g (g = 0, 1), θ = Latent ability status for an individual test-taker, a_j = Discrimination parameter for Item j, which can be allowed to vary across groups, $b_{j,g}$ = Difficulty (location) parameter for Item j in Group g, and $d_{i,j,g}$ = Threshold parameter for Category i of Item j in Group g. The "group step difficulty," $b_{i,j,g}$, for a given Category i can be obtained as $b_{j,g} - d_{i,j,g}$. We intended to model any groupwise differences in Item j discrimination by allowing the category threshold parameters for an item to vary across groups.

We estimated the difference between generalized partial credit item difficulty parameters in test-taker groups that had (1) and had not (0) experienced relevant topical instruction, BDIFF, as an instructional sensitivity index (Equation 2):

$$BDIFF = b_{j,0} - b_{j,1} \qquad (2)$$

with positive values indicating IS favoring the group with more opportunity to learn, as expected in theory, but negative values also possible.

In our formulation, all items had a multiple-group model, except one scale anchor item that had parameters held constant across the groups, a "free baseline" strategy (e.g., Nye & Drasgow, 2011). The mean latent science achievement score for each opportunity-to-learn (OTL) group was constrained to 0, while that for the No OTL group was freely estimated. This parameterization allowed us to distinguish between group differences in item difficulty that could be attributable topic-specific opportunity to learn, and differences in overall science achievement. Because TIMSS is an observational study, the student groups with and without opportunity to learn about the broad content topic of each item presumably differed on other unmeasured characteristics, which raises questions about comparing their item performance in an IS index. However, previous studies of mathematics item instructional sensitivity demonstrate that students' current total achievement scores are the strongest predictor of their performance on individual test items, and that adjusting for student OTL group differences in other background variables seldom changes the conclusions about an item's instructional sensitivity (Li et al., 2017; Muthén et al., 1991).

In previous work (Traynor et al., 2025), we used Monte Carlo simulation of responses to a short test composed of ordinal items that followed the generalized partial credit model to evaluate the performance of the BDIFF item instructional sensitivity index under "questionable" use conditions. Across all conditions, we found a systematic small positive bias in the estimated value of BDIFF. Percentage bias tended to decrease with combined group sample size, and was generally less than 2% in the conditions with 500 test-takers in the opportunity and no opportunity to learn groups. Overall, we concluded from the simulation that the BDIFF index seems to perform well enough that it may be useful to evaluate test item instructional sensitivity when several hundred test-takers have completed each item.

To address Research Question 1, we computed the IS effect size index for the 32 constructed-response science achievement items, each across two years. Missing data was handled by using full-information maximum likelihood estimation with robust, classroom cluster-adjusted standard errors. The criterion that we used to evaluate the stability of the IS index results for the same item across years was whether decisions about the index's relative magnitude (e.g., "moderate," "high") based on Roussos et al.'s (1999) criteria change. We also report a delta-method 90% confidence interval for each index value.

3. Results: Research Question 1

We applied each two-group generalized partial credit model to students' responses from the TIMSS Grade 8 English-version science constructed-response item sets that were administered in 2011, 2015 and 2019, and computed the BDIFF step difficulty instructional sensitivity index for each adjacent score category boundary (0,1 and 1,2 if applicable). Originally, we conceptualized BDIFF as the difference in overall item difficulty location for the OTL and No OTL student groups. As shown in Table 4, when items had more than two score categories, we found that the magnitude of BDIFF values sometimes varied appreciably across the category boundaries within an item, as well as across items. Overall, we found that the constructed-response science item scores had varying instructional sensitivity, consistent with earlier results about mathematics items (Li et al., 2017; Muthén et al., 1991). Twenty-nine out of 37 item score categories had consistent instructional sensitivity effect size classifications across years: 16 were negligible, 2 were moderate positive, and 11 were large positive. Eight item score categories that had varying instructional sensitivity effect size classifications; six of those changed to an adjacent classification category (e.g., negligible to moderate positive). Only two items—42228C and 42402—had large positive IS in one year and negligible IS in the other testing year. Many items that showed large IS had specific content topics about matter or energy (e.g., Classification of Matter; Morell et al., 2017). Recall that item IS effect sizes have been shown to decrease as an item's solution process requires near or far transfer of content knowledge that students learned previously (Ruiz-Primo et al., 2012; see also Section 1.3).

Thus, moderate positive and even ‘negligible-to-moderate positive’ IS effect sizes may be evidence that a test item can measure learning.

Regarding our Research Question 1, we found that the instructional sensitivity index values for a particular item tend to vary considerably across testing years. However, the conclusions about each item’s degree of instructional sensitivity (i.e., negative, negligible, moderate positive, large positive) when the index magnitude is compared to Roussos et al.’s (1999) effect size criteria are quite stable, changing across the 2011/2015 testing cycles for only three of the items (42088, 42228C, and 42402). The evidence suggests that instructional sensitivity is a reasonably stable property of test items, so information about item IS could be useful during future test item writing and/or form assembly.

We then computed the BDIFF item instructional sensitivity index for TIMSS Grade 8 science constructed-response item sets that were adapted from the Arabic and English source versions in 2011. As shown in Table 5, eight of the 32 items had index values that differed widely, depending on whether students’ item response data came from the Arabic or English source version. The conclusions about each item’s degree of instructional sensitivity when the index magnitude was compared to Roussos et al.’s (1999) effect size criteria were reasonably stable for the other 24 items, though. Returning to Research Question 1, it appears that item IS index values have some generalizability across the test language versions, but perhaps not enough to use IS index results from one language version to inform test item writing practices for other language versions.

4. Methods: Research Question 2

After our team identified 20 constructed-response science test items and rubrics that yielded scores with either low or high instructional sensitivity from an earlier study (Traynor et al., 2025; see Table 6 and Figure 1), eight master science teachers from Botswana, Chinese Taipei, Türkiye, and the United States analyzed the items’ content features. These four education systems’ TIMSS assessment items were all adapted (Botswana, United States) or translated (Chinese Taipei, Türkiye) from the English source version. Each educator conducted item content analysis facilitated by one co-investigator, following a semi-structured interview protocol. The study procedures were designed by the co-investigator in each research location. The research questions remained consistent, but the procedures (e.g., recruitment method, language of communication, interview location type, interview protocol, participant compensation) were adapted for each national education system. All teacher participants gave signed informed consent to participate, after reading information about the study and discussing it with the local co-investigator, consistent with each jurisdiction’s regulatory requirements.

4.1. Participants and Contexts

Eight master science teachers, two from each national education system, who fit our participation criteria were chosen as study participants using purposive sampling. The master science teachers were current teachers who had at least 5 years of lower secondary science teaching experience and held the highest educational credential typical for Grade 8 (e.g., Form 2) science educators in each education system (see Table 7). These criteria assured that our participants had the expertise and experience to be able to thoroughly evaluate the content of Grade 8 science test items and scoring rubrics. They were recruited and consented specifically for participation in this study; four of them were previously known to the investigators from existing research relationships.

4.2. Interview Protocols

In the earlier quantitative study phase, a student's opportunity to learn about the broad topic of each test item was measured by their science teachers' yes or no responses on the TIMSS teacher questionnaire. We maintained a similar definition when we interviewed teachers about the test item content. Specifically, we designed our interview questions to compare the predicted cognitive response processes of students who did, and did not, have any school instruction about each item's broad topic (e.g., Chemical Change or Reproduction of Life; Martin & Mullis, 2012). Each item content analysis interview had two parts. During the first segment, each teacher was presented with 20 test items and their corresponding scoring rubrics, one at a time, and asked to predict students' response processes. In the second segment, we asked teachers to determine common aspects of students' predicted cognitive response processes for four items that had high (BDIFF $\geq .30$) and four items that had near-zero (BDIFF $\leq |.15|$) instructional sensitivity index values, spanning content topics. At the end of each interview, the interviewer explained the concept of item IS statistic values from comparing opportunity-to-learn student groups, and shared a summary of the results from the earlier quantitative study, so participating teachers could recognize how their content analysis would contribute to test development guidance.

The original interview protocol was trialed by one master science teacher. Given the teacher's input, we assessed whether the questions would prompt unambiguous responses that were relevant to our research questions, and overall usability of the protocol. Considering that educator's experience, we reduced the target number of items to be reviewed by each teacher from 24 to 20. We believed that patterns in teachers' judgments about the item and rubric content would be detectable if they evaluated multiple items from each content area with high, and with negligible, instructional sensitivity. Cognitive interviews typically reveal patterns in respondents' thought processes after analyzing a relatively small number of items, especially when they have relevant professional expertise (Willis, 2015). We also planned the item presentation order so that items with extreme instructional sensitivity index values, either high positive or near-zero, were presented first. The item order was staggered by IS index magnitude and by content area: Biology, Earth Science, Chemistry, and Physical Science. This strategy produced the sharpest contrast in the results that was possible, if any teachers reached the maximum interview time before reviewing all the items.

Appendix 1 presents the protocol that was adapted in each study location, and used for semi-structured item content analysis interviews. The role of the interviewer was as a facilitator of each teacher's item content analysis. Once a teacher became familiar with the interview structure, they were encouraged to lead the analysis of the items and rubrics. The interview questions asked teachers to predict: (i) how eighth-grade students would interpret each item and their response strategies. Then, while also viewing the rubric, the teachers were asked to predict: (ii) how students might use content that they learned in school to produce partially- and fully-correct responses, and (iii) if there were processes by which students without previous formal science instruction about that item's content topic were likely to produce partially- or fully-correct responses. The interviewers emphasized that we were not interested in situations of "teaching to the test items," but rather regular content-area instruction about certain broad topics. Teachers generally spoke at length about the content features of each item with little prompting, but interviewers occasionally prompted the educators to elaborate their predictions about students' cognitive response processes or their rationale. We anticipated that the reliability of teachers' predictions about students' cognitive processes would be modest, as has been observed for other content expert raters (Hambleton & Jirka, 2006), but that their content analysis would still be useful for identifying causes of negligible and high item instructional sensitivity.

Teachers were interviewed individually in a public location that they chose as adequately private (e.g., a library) or at their school campus, consistent with local research norms. The item content analysis interviews occurred between May and August 2024. The interviews were audio- or videotaped, and interviewers recorded notes about teachers' predictions. Each teacher's interview was 2 to 4 hours in duration, across one or two sessions.

4.3. Thematic Analysis Methods

Following their teacher interviews, each co-investigator independently began reflexive thematic analysis of a transcript of the conversational exchange (Braun & Clarke, 2022; Madill & Gough, 2012; Terry et al., 2017). Our theoretical orientation toward teachers' statements about the test item content and their predictions about student cognition was critical realist; we interpreted teachers' statements as providing indirect information about a true local reality, from the perspective of their own context and experiences.

4.3.1. Interview Transcript Text Coding. Verbatim interview transcripts from both teachers in each national education system were coded by the local co-investigator. Each coder read each interview transcript twice fully before commencing to code it. Since our research question addressed the test item content, the unit of content analysis was the transcript per item. Coders assigned initial codes (i.e., short descriptive labels) to transcript text that seemed relevant to Research Question 2. For example, some codes were "giving explanation beyond short answer may be a major challenge for some students," "common in daily life experience", and "students' understanding of the technical terms may not be precise." Coders assigned text codes to both teachers' transcripts for a particular item, then repeated this process for the next item. This process allowed the coder to understand individual teachers' predictions about students' approaches to a test item as completely as possible, because occasionally one teacher's item content analysis provided context information that clarified the other teacher's assertions, and prompted revision of a code in the other teacher's transcript. It also allowed the coder to identify salient features of each test item that reflected the teachers' collective professional judgment, rather than opinions that were unique to one person. Relying on their collective judgment contributed to the trustworthiness of our results for our purpose of informing test development guidelines in general, beyond any single teacher's classroom.

Our initial coding of the transcripts was primarily inductive from the data, rather than deductive from theory, because there was little existing evidence about test item features that affect item score instructional sensitivity. However, a few codes were informed by the earlier research findings outlined in Section 1.1 of this report. We applied primarily direct semantic, rather than latent, coding, unless we had high confidence in the implicit meaning of a teacher's statement. During the first round of coding, each coder shared their initial codes with another investigator on the team for feedback. After completing the first round of coding, which included iterative revision of the codes within and across items, each coder re-read their transcripts fully once more to finalize the text codes that would be brought to the cross-national theme development meetings.

4.3.2. Theme Development. The co-investigators exchanged our transcript codes, approximately 15 to 25 per research site, and met to identify provisional themes from the item content analysis interviews. An important advance during this phase was that we realized teachers' analysis of the items' scoring rubrics was not only consequential for their overall evaluation of each item, but also for test development guidelines to a greater extent than we had anticipated. Although our original research question asked about the "content features" of "items," we updated it to refer to "item/rubric pairs." Considering the research question, we grouped the transcript codes that expressed similar ideas about test item or rubric features, initially attaching some codes to multiple themes. A few codes were adjusted to be more general, so they could potentially be applied to additional items. Transcript codes that were not closely connected to any provisional theme were left aside in this phase. After the meeting, the coders were allowed but not required to revise their codes.

To bridge our thematic analysis back to the item instructional sensitivity effect size index values, during our next meeting, we re-arranged the transcript codes into two groups: codes from items with negligible, and with high, instructional sensitivity. Then, we grouped the transcript codes into themes within each item set. Because the study purpose was to inform test item writing guidelines that would be useful across countries, as much as possible, we prioritized transcript codes that appeared across countries to define each theme. We did not restrict overlap of the themes across the two item sets, but we did limit the themes to concepts that seemed pertinent to future test item or rubric development activities. A theme had to be supported by codes from multiple items in multiple countries and imply a distinct recommendation for test item writing guidelines. Relying on the replication of text codes, as well as the consensus judgment of investigators who were deeply familiar with each national education system, contributed to the trustworthiness of our results for our purpose of informing test development guidelines in general. When provisional themes were specific to one national education system, we report them as noteworthy results, rather than as themes.

To summarize our findings without eliminating details that might be useful for test development, we demoted a few provisional themes as 'sub-themes' of another theme. Finally, we used quotation extracts that were linked to the priority code(s) to name the themes from the teachers' item content analysis interviews. The quotation extracts were lightly cleaned prior to reporting, removing utterances that were non-informational for our study purpose (e.g., filler words).

5. Results: Research Question 2

This section provides a comprehensive analysis of the findings from teacher interviews conducted across four countries: Botswana, Chinese Taipei, Türkiye, and the United States. Eight teachers participated in this study, and their insights were carefully examined through thematic analysis to uncover key themes related to students' science education and assessment. The analysis focused on understanding the relationship between the test items, the intended and enacted curricula, and students' real-life experiences. Teachers' perceptions of student conceptions, cognitive processes, and the effectiveness of instructional strategies were central to this exploration.

The themes identified in this section reflect common patterns and differences observed in teaching practices and students' understanding across the four countries. Each theme is supported by detailed quotes from teachers, which are attributed using teacher codes for anonymity. These quotes not only highlight the unique educational contexts but also provide a deeper understanding of how students' real-life experiences and prior knowledge influence their performance on science assessments.

Furthermore, the thematic analysis sheds light on how different instructional approaches and curriculum designs affect student outcomes, offering valuable insights for improving both teaching and assessment practices in science education. The cross-national comparison presented in this section highlights the shared challenges and opportunities faced by educators in fostering students' scientific literacy and critical thinking skills.

Theme 1: Impact of Curriculum Standards on Student Performance

The relationship between curriculum standards, test content, and student performance emerged as a significant theme. Student performance is heavily influenced by the relationship between the curriculum and the test items. This includes vertical sequencing of curricular standards, factual knowledge retention, and the rigor of standards in each country. Examples of initial transcript text codes that supported this theme were: *correct answers will be influenced by understanding science terms* and *retention of factual knowledge depends on grade-level sequence in curricular standards*.

- **Students' Item Responses Rely on Learning from the Implemented Curriculum**
  - *T2 (Botswana)*: "Most students may not be conversant about some elements and compounds; they may just get correct some areas by guesswork."
  - *T4 (Chinese Taipei)*: "This is something covered in Chinese Taipei's elementary science curriculum, where students observe that water expands when frozen. In middle school, this concept is explained in more detail, relating it to 'density.'"
  - *T6 (Türkiye):* "They don't have prior knowledge about wasp venom being basic. They might have learned it from alternative sources, but it's not part of the curriculum."
  - T8 *(USA)*: "I don't know if my eighth graders would get this right, because we don't teach simple circuits anymore. This isn't in our state standards, but it might be covered elsewhere."
- **Performance Depends on the Vertical Sequence of Curricular Standards, which Affects Information Retention**

Teachers emphasized how the sequence in which topics are introduced across different grade levels influences student performance. Teachers from multiple countries pointed out that students' recall can be reduced due to instructional gaps or limited repetition practice in the curriculum, particularly when students are expected to connect knowledge from different grade levels.

  - *T1 (Botswana)*: "Our students sometimes struggle with topics that were introduced two years ago, especially when they were not reinforced later."
  - *T2 (Botswana)*: "Students may not understand topics like mixtures because they haven't been introduced to the concept in class yet."
  - *T3 (Chinese Taipei)*: "Students might confuse different forms of sugar (e.g., brown sugar, which is a mixture), and this can cause issues when determining whether it is a compound or mixture."
  - *T5 (Türkiye):* "Students in eighth grade sometimes forget what they learned in sixth grade because the curriculum doesn't reinforce it enough."

- T7 (USA): “By eighth grade, students might not remember concepts we taught them in sixth grade, like the speed of light versus sound. If it's not part of the eighth-grade standards, they may not be prepared to answer those questions.”

The first theme explored how the vertical sequencing of curriculum standards across grade levels significantly influences students’ memory retention and performance. Teachers reported that students struggle with concepts introduced in earlier grades if these concepts are not regularly reinforced. Teachers from all countries observed that the rigor and content of curriculum standards directly affect students’ performance.

Theme 2: Connections among Instruction, Test Content, and Students’ Responses

This theme explores the connection between classroom instruction (enacted curriculum) and the content of test items, as well as how students leverage everyday knowledge and experiences to progress toward expert levels of science understanding. Examples of initial transcript text codes that supported this theme were: *performance can be influenced by teacher use of correct jargon and emphasis* and *difference from context used during teaching*.

- T1 *(Botswana)*: “Without learning at school, they won’t be able to explain why the bottle cracked. They need content for this.”
- T2 *(Botswana)*: "Students may not understand topics like mixtures because they haven’t been introduced to the concept in class yet”. Similarly, T2 noted that “They think elements like helium and gold are unfamiliar because they haven’t learned about them yet.”
- T3 *(Chinese Taipei)*: “Because the question states that wasp venom is a base, students will immediately associate it with acid-base neutralization, a concept they’ve learned in class.”
- T4 *(Chinese Taipei)*: “The river question is straightforward since students already know from geography class that rivers flow downhill. However, they may make mistakes when it comes to drawing the arrow correctly.”
- T5 *(Türkiye)*: “Our students tend to do better on items that mirror the examples I use in class. When the wording or scenario is different, they struggle.”
- T6 *(Türkiye)*: “In sixth grade, we teach renewable and non-renewable energy sources and their impact on sustainability. We also cover matter cycles in eighth grade.”

This theme highlighted the strong connection between classroom instruction and student performance on assessments. Teachers noted that students perform better on test items that align closely with classroom examples. When test content departs from these familiar contexts, students may struggle to transfer knowledge that they have learned, particularly when faced with unfamiliar vocabulary or scenarios. Teachers also stressed that everyday knowledge, although useful, can sometimes hinder students when it is not aligned with scientific reasoning taught in class.

Theme 3: Everyday Science Knowledge and Naïve Conceptions

This theme focuses on naïve conceptions and the cognitive strategies students use to process test items, highlighting how teachers perceive these challenges across different countries. Teachers from all countries discussed how students' everyday experiences shape their understanding of scientific concepts. In some cases, everyday knowledge supports correct answers, while in others, it can lead to oversimplifications.

- **Common Naïve Conceptions of Science Phenomena**

Across the countries, teachers noted frequent naïve misconceptions that hindered student test performance, often rooted in everyday experiences. Teachers also emphasized the importance of scientific vocabulary for understanding and correctly answering scientific tasks. Students often struggle when they lack the necessary terms, even if they understand the concept.

  - *T3 (Chinese Taipei)*: "Students might say that lightning is faster than thunder based on their life experiences, but they might not connect it to the scientific concepts of light and sound speed."
  - *T4 (Chinese Taipei)*: "If students haven't learned about buoyancy, they may struggle to answer correctly, even if they are familiar with the concept of floating from daily life."
  - *T5 (Türkiye)*: "They often confuse different scientific terms, like energy and power, feather and hair, or using them interchangeably."
  - *T6 (Türkiye)* discussed a similar challenge where students might understand the basic processes of a scientific phenomenon but struggle to express it accurately: "They might know the concept, but without the correct vocabulary, they can't explain photosynthesis properly, or they know wasp venom is basic, but they can say that it won't neutralize."
  - *T8 (USA)* similarly stated, "They wouldn't say maybe recessive and dominant, but they would understand that there was mating or breeding involved."
  - *T8 (USA)*: "A lot of my students mix up mass and weight, thinking they are the same. It's a hard misconception to break."

- **Naïve Conceptions Stemming from Everyday Experience**

Teachers from all four countries noted that students frequently rely on their everyday life experience when answering questions, even when that knowledge does not fully align with the scientific concepts being assessed.

- *T1 (Botswana)*: "Students fail to understand the concept of weight and mass. In everyday life, we speak of weight using units of mass, which confuses them."
- *T2 (Botswana)*: "They might base their explanation on everyday experiences, because at home they see a cow giving birth to a calf that looks like it, the same with cats, donkeys even people."
- *T3 (Chinese Taipei)*: "In life, students know that adding water dilutes substances, which is why they might answer the question about concentration correctly. But they may not fully understand the scientific reasoning behind it."
- *T5 (Türkiye)*: highlighted a similar issue where students' real-life experiences can lead to incorrect assumptions: "Students sometimes think that just because they see a flower grow in sunlight, they assume light is the only thing necessary, ignoring water or nutrients."
- *T7 (USA)*: noted that relying on story narratives can sometimes introduce naïve conceptions: "There's also some, like mythology almost that kids carry with them... About storms and lightning and thunder."

- **Reliance on Everyday Experiences to Respond to the Test Items**

Teachers observed that students tend to rely on their everyday experiences to respond to test items, even when they have had formal instruction about the item's content topic, which can lead to both partially correct and incorrect answers.

- *T2 (Botswana)*: "I think, as for water, most of the students, they will write carbon dioxide, but with this one, yeah, water, I think why am I saying this is because even from the younger age, students know that if you have a small garden at home definitely, your plants need water. If you go for some days without, then the plants die. So, I think the everyday life experience matters."
- *T4 (Chinese Taipei)*: "Students may have some exposure to concepts like baking soda or air pollution, but without the scientific foundation from class, their answers might be incomplete or incorrect…In terms of life experiences, some things are more familiar, like sugar, iron, and air…Sodium bicarbonate is baking soda, and students might not answer this because they don't know."
- *T7 (USA)*: "Even students with life experience observing rivers may just say 'rivers flow to the ocean' without providing a full explanation."

- *T8 (USA)* described how students rely on everyday experiences to make sense of natural phenomena: “Water is gonna go from high to low, based on just general observations.”

Conceptions rooted in everyday experiences emerged as a critical challenge for students across all countries. Teachers emphasized that students frequently rely on non-scientific terminology or oversimplified reasoning, which leads to incorrect answers. These misconceptions often stem from real-life experiences that conflict with scientific principles. For example, students confuse concepts like mass and weight due to everyday language, or they misinterpret scientific terms such as "base" and "neutralization" because of limited exposure to scientific vocabulary.

Theme 4: Specific Life Experiences as a Factor in Performance

Certain life experiences—whether related to the student’s local environment or family background—were noted as giving some students an advantage on particular test items.

- **Influence of Environmental Exposure on Contextual Knowledge and Test Performance**

Teachers frequently mentioned that students who live in particular environments, such as rural areas or regions with specific climatic conditions, have an edge when answering relevant questions. Teachers from different regions noted how cultural factors can shape students' understanding of scientific concepts, sometimes leading to inequalities in performance on certain tasks.

- *T1 (Botswana)* mentioned about students in rural areas who understand plant life cycles better due to their exposure to farming: "Students who have been involved in farming understand concepts like germination better."
- *T2 (Botswana)* added that students' experiences with their environment affect their interpretations: “Students who live in areas where there’s frequent rain might answer differently about the water cycle than those in dry regions.”
- *T2 (Botswana)* added that students' experiences with their environment affect their interpretations: “Students who live in areas where there’s frequent rain might answer differently about the water cycle than those in dry regions.”
- *T3 (Chinese Taipei)*: “It depends on whether students have experience with planting in fields or outdoor observations. In Taiwan, this varies significantly. In urban schools, opportunities might be fewer…Students with more exposure to nature may perform better on certain questions.”
- *T5 (Türkiye)*: "Those who have experienced natural disasters, like floods or earthquakes, tend to answer questions on these topics more accurately."

- *T6 (Türkiye)* also highlighted that students living in rural areas might better understand certain concepts related to nature and agriculture: "Those with a farming background can often explain processes like plant growth without much difficulty."
- *T8 (USA)* mentioned that students with particular life experiences, such as farming or hunting, are more likely to give correct answers related to those contexts: "Students who come from farming backgrounds might have a better understanding of selective breeding."
- *T7 (USA):* "Students that have studied photosynthesis outside of school, or even observed it in a garden, might get this question right."
- *T8 (USA)* observed that students who live in areas prone to natural disasters like earthquakes or tornadoes are more likely to understand certain scientific phenomena: "I think somebody who grew up in an area like California, Hawaii, or anywhere in the Ring of Fire is going to say volcano, or fault line without much problem."

▪ **Advantages of Familial Background**

Life experiences within families also play a role in how students approach science questions.

- *T1 (Botswana)* noted that students with agricultural backgrounds tend to perform better on questions about plant life cycles due to their familiarity with farming practices: "Students who have been involved in farming understand concepts like germination better."
- *T2 (Botswana)*: "Those with gardens at home will know that plants need water, so they will get the answer correct even if they haven't fully learned the topic."
- *T3 (Chinese Taipei)*: "Students who have done experiments with freezing liquids at home, like freezing Yakult, will have a better understanding of how water expands when frozen."
- *T5 (Türkiye)* stated, "Students who live in rural areas can recognize this situation" referring to an item about plant roots and soil layers.
- *T6 (Türkiye)*: "Students with parents working in agriculture or engineering show a better understanding of certain items."
- *T8 (USA)*: "Students whose parents are in scientific fields or farming have better insights into certain questions."

Teachers identified specific life experiences as a significant factor in how students approach and perform on science assessments. Students from rural or agricultural backgrounds often perform better on items related to plant life cycles or environmental processes due to their direct exposure to these concepts in their daily lives. Conversely, students without such experiences may struggle, highlighting the importance of contextual knowledge in scientific understanding. Teachers across all four countries emphasized that environmental exposure can either enhance or limit students' ability to engage with the test item content.

Theme 5: Student Understanding of the Item, and Impact of the Scoring Rubric on Performance

This theme explores teacher perspectives on the clarity of test items and the fairness of scoring rubrics, focusing on how these factors can impact student performance and their ability to demonstrate understanding.

- **Clarity and Complexity of Test Items**

Teachers, particularly in Chinese Taipei, Türkiye, and USA, raised concerns about the clarity of some test items. They noted that certain questions, whether due to their complexity or the way they were presented, often led to student confusion, preventing them from fully engaging with the scientific concepts being tested.

- *T1 (Botswana)*: "Some students will fail to interpret the diagram correctly because they expect the mountain to be at the top and the ocean at the bottom."
- *T4 (Chinese Taipei)*: "The question about river flow could confuse students, even though the diagram should make it clear that rivers flow from high to low. Some students may still struggle with drawing the arrow correctly."
- T5 (Türkiye) explained, "These topics are covered in higher-level classes. It could be very challenging for students at this level," referring to an item on Ohm's Law that required mathematical calculations students had not yet learned.
- *T6 (Türkiye)* similarly emphasized that lack of clarity in the item content can cause misinterpretations: "Certain diagrams or instructions are not very clear, which leads to misinterpretation of the questions." This confusion can lead students to provide answers based on irrelevant details, even when they understand the underlying science.
- *T7 (USA):* "I could see some students thinking that water might neutralize a base, and their error would be based on missing that key scientific vocabulary."
- *T8 (USA)* highlighted that students often struggle with overly complicated items: "I can see students struggling with this because I'm looking at her going, 'Well, it's floating' in all three diagrams. It's just, I don't know how much they're gonna pay

attention to the differences in density." This illustrates how an item format may impede students' interpretation.

- **Rubric Criteria for Partial Credit**

Teachers also questioned some of the scoring rubric criteria, particularly for situations when students provided answers that were partially correct but lacked specific scientific terminology. They felt that certain rubric performance level descriptions failed to capture students' legitimate partial understanding. These teachers emphasized the importance of recognizing student effort and conceptual understanding in the rubric criteria.

- *T1 (Botswana):* "They may get the first part of the answer right, but their explanation will be too shallow to receive full credit."
- *T2 (Botswana):* "They might get partial credit for their answer, but without the correct scientific terms, they will struggle."
- *T3 (Chinese Taipei): "*Most students would respond based on height differences, but they wouldn't explain that water can't flow uphill. The rubric expects a deeper explanation." *T3* also added that "The scoring rubric seems quite clear for questions where students just need to check the correct answer, but partial credit should be awarded for effort."
- *T4 (Chinese Taipei):* "Students may understand the concept of buoyancy but fail to explain it in scientific terms. They might lose points even though their reasoning is correct."
- *T5 (Türkiye):* "This is a higher-level question on buoyancy that requires multiple levels of explanation, but the rubric demands more reasoning than students can provide."
- *T6 (Türkiye):* "Student give the correct answers but lose points due to a lack of terminology."
- *T8 (USA):* "Without scientific vocabulary, students might still get the general idea, like understanding that deeper roots provide stability, but they might not use the correct terms."

The final theme focused on the clarity of test items and the performance level descriptions in the scoring rubrics. Teachers expressed concern about test items that seemed likely to be misunderstood by students, which would detract from their ability to demonstrate their knowledge. There was variation across national education systems in the amount of test item elements that the teachers believed to be unclear. Additionally, teachers noted that the rubrics sometimes failed to account for students' partial understanding, particularly when they lacked only precise scientific terminology. They noted that such disconnects between students' conceptual understanding and their ability to express it using the expected language could result in lower item scores, despite a correct understanding of the scientific principles involved.

Overall, the thematic analysis of master science teachers' item content analysis interviews revealed insights into the relationships among science curricula, instructional practices, and student performance in science assessments. The findings were organized around five major themes, each addressing the complexities and challenges in aligning test items with curricula, the role of instructional context, the impact of students' science conceptions from everyday life experience, and the effect of assessment design on student performance. This thematic analysis highlighted shared challenges in science assessment and instruction across Botswana, Chinese Taipei, Türkiye, and the United States, including cross-national variation in the intended and implemented curriculum, the influence of daily life experiences on students' understanding of natural phenomena, and the adhesion of naïve conceptions. These findings offer insights for science test item design that can effectively capture students' conceptual understanding and cognitive processes.

6. Study Limitations

A notable limitation of our study design is that enlisting a fixed number of master science teachers did not necessarily ensure thematic saturation. Although we recruited science teachers from four national education systems, and structured our recruitment criteria so the two participants invited in each location would differ on at least one characteristic that we hypothesized as salient to their perceptions of the test items, two participants can be only partially representative of teacher perceptions in an education system. We did not observe new themes or many novel codes emerging from the item analysis of the second teacher in each country, suggesting that the data might be approaching saturation. However, with no means to recruit additional science teachers beyond the first eight, we could not confirm thematic saturation, which limits the generalizability of our findings. Another limitation is that although item instructional sensitivity has some stability across test language versions, our first study phase showed that the IS differences across language versions can be pronounced for certain items. Accordingly, we acknowledge that the science test item set content analyzed by the master science teachers might better be viewed as four item sets, containing items that had been adapted or translated to suit use in each national education system, with items from the Botswana and United States versions perhaps being most similar. Finally, although we relied on multiple data sources to draw conclusions—statistical analysis of students' item response data and teachers' predictions about the item content—we did not attempt to triangulate the findings with additional data sources, for instance, cognitive interviews of eighth-grade students taking the test items. This might be a useful avenue for further research.

7. Conclusions

We presented a theoretical framework about the value of achievement test item score instructional sensitivity when students' opportunity to learn is measured as broad content topic instruction, rather than instruction about exact item content or format. Items with moderate or high IS seem desirable from a validity perspective if the test scores will be interpreted as measuring curricular achievement (that is, learning in school). In the context of international large-scale assessment, items with moderate or high IS also seem acceptable from a fairness perspective if most or all participating countries' implemented curricula introduce those content topics. Then, we evaluated evidence about test item score IS in a diverse, multi-national test-taker population over time, which allowed us to establish that the IS index magnitude of any particular item tends to be fairly stable, rather than an unpredictable result of test-takers' interaction with the item on one testing occasion. Attempting to extend the generalizability of item IS indices to multiple language versions of the same test, we found that one-fourth of the items had index values that differed widely depending on whether students' item response data came from the Arabic or English source version. (IS index magnitudes of the other items changed relatively little across the test language versions.) We conclude that item instructional sensitivity is a stable property of test items when the test language and scoring rubric remain consistent, so information about item IS could be useful during future test item writing and form assembly.

Given our results from the first study phase, we then pursued structured content analysis of test items with negligible and high instructional sensitivity. Our purpose was to inform item writing guidelines when item instructional sensitivity is necessary validity evidence for a planned test score use (e.g., when the test scores will be interpreted as measuring students' learning in school). After we identified 20 constructed-response science test items and rubrics that yielded scores with either low or high instructional sensitivity, eight master science teachers from Botswana, Chinese Taipei, Türkiye, and the United States analyzed the items' content features. The science teachers anticipated students' interpretations of and response processes for each test item, as well as students' likely performance given the scoring rubric. As we invited the teachers to talk about the items' content in the context of their implemented curriculum, they revealed further insights about students' common reactions to science test items.

Considering the themes that emerged from the item content analysis, we summarize item and rubric content features that co-occurred with high or negligible item instructional sensitivity in Table 8. The master teachers explained how these test design features interact with students' content exposure in their science classes to affect item instructional sensitivity. In the remainder of this report, we discuss the implications of our findings for test item writing and rubric design.

7.1. Test Instructions Can Prompt Scientific Explanation and Science Vocabulary Use

Items about academic science concepts often had scores with high instructional sensitivity, particularly when the rubric score categories required students to explain or elaborate their responses. Educational test item writing guidelines emphasize aligning the item content with curriculum objectives (Albano & Rodriguez, 2018), so we anticipated that academic science item scores would show instructional sensitivity. However, teachers from all four national education systems described students who would struggle to explain, or would not recognize the extent of explanation that was expected (Rottman & Keil, 2011).

Scientific explanation is particularly challenging to teach students because it requires metacognitive regulation (Özdem-Yilmaz et al., 2017; Wang, 2015). Participants in this study perceived that some students would provide insufficient information when explaining a science phenomenon, or neglect to use evidence from the item to support their response (Braaten & Windschitl, 2011; Samarapungavan, 2018). They worried that students might substitute vague, informal explanations for the scientific accounts that they had learned in school. This challenge may be particularly pronounced if students are unfamiliar with the item prompt structure, or misinterpret the degree of elaboration required to receive full credit.

Recommendations: Allowing eighth-grade students to respond in multiple formats (e.g., drawing, diagramming, or mapping) could provide alternative ways for students to express their scientific understanding rather than being limited to written responses. It is possible that the current computerized version of the TIMSS assessment permits flexibility of this nature.

The test instructions offer an opportunity for the test developer to elicit higher-quality responses from students, so should be carefully evaluated (Lakin, 2014). If some students could explain—but do not recognize that explanation is required, and attain a lower score, then this would affect validity. The instructions at the start advised students to “answer as completely as possible,” which may prompt them to respond with reasoning. Emphasizing this direction, perhaps by repeating it or making the instructions more concise, could convey its importance in the assessment of students’ responses. Stating that students should use ‘science ideas’ in their explanations could encourage students to emphasize science content knowledge in their responses. (In the TIMSS context, implementing this would require separating the mathematics and science instructions.) The test instructions also present a student’s constructed response to one item as an example, but the response is a short sentence, which may not convey to students that they should attempt to elaborate. Presenting an example item response that demonstrates a well-elaborated explanation may further clarify expectations could help students perform to the best of their ability.

7.2. Monitoring Use of Science Vocabulary in the Item Text or Scoring Rubric Criteria

Science vocabulary allows phenomena to be described with precision and can facilitate communication with other scientists, but certain words are difficult to learn (Brown & Concannon, 2016; Cervetti et al., 2015). We found that science vocabulary use in either an item or its scoring rubric criteria was associated with high item IS. The vocabulary content presumably elevated IS by increasing the item's difficulty for students who had not experienced relevant instruction.

Our master science teachers were particularly concerned that students would not know which rubric criteria called for responses containing science vocabulary. They also noted that terminology within item texts could sometimes take on different meanings when interpreted through students' prior everyday experiences. For instance, learners may lack the command of academic science language to distinguish the terminology that appears in the test items (e.g., '*hair*' versus '*feather*' as a colloquialism in Türkiye; '*sucrose*' versus '*brown sugar*' in Chinese Taipei's living context). Teachers observed that other students might recognize the scientific concepts being assessed but struggle to interpret the terminology. Intentional instructional strategies by educators can teach students to "question, synthesize, visualize, and represent…vocabulary while they read" (Brown & Concannon, 2016, p. 404), but students' interpretations of the items will still be shaped by their experiences in their culture (e.g., Aikenhead & Jegede, 1999).

Recommendations: To ensure that conceptual understanding receives sufficient weight in the test scores, careful attention should be given to the proportion of test items and rubric performance levels that require the use of specialized scientific vocabulary. During item development, it is important to evaluate whether each vocabulary term has a close and accurate translation in different test language versions and whether the translated term is commonly used in instructional settings (International Test Commission, 2019). Systematic piloting of the items with diverse student populations before operational use can also allow test developers to examine how students interact with certain scientific terms and whether they create unintended barriers to comprehension. Based on the findings, refinements can be made to ensure that the vocabulary choices align with students' instructional exposure, while preserving the scientific rigor of the test content.

7.3. Characterizing Partially-Correct Science Reasoning in the Scoring Rubric Criteria

Scientific thinking involves not only the recall of factual knowledge, but also the ability to apply, evaluate, and articulate concepts effectively. While limited scientific reasoning and insufficiently developed explanations arose as themes in our item content analysis interviews, the master science teachers also emphasized that the evaluative criteria of a particular scoring rubric constrain the score outcomes that are possible for a student (e.g., Brookhart, 2018). The teachers often predicted that their students could provide correct answers to an item. However, after reviewing the corresponding scoring rubric, they perceived many students were then unlikely to be assigned any points, because the performance level description for the 'partially correct' category limited the variety of responses that could be scored in that category. For example, a student might correctly describe the purpose of a biological process, but lack complete articulation of its mechanisms.

Recommendations: Intentional design decisions should be made about the balance among science concepts, explanation, academic science vocabulary, and everyday science knowledge in any 'partially correct' rubric level, because it can affect the item's instructional sensitivity. Our master science teachers recommended that more of the item scoring rubrics should incorporate intermediate performance level descriptors that reflect accurate concepts or logical reasoning that is correct but incomplete. Including additional exemplar student responses in the intermediate performance descriptors, as well, may facilitate reliable scoring. A rubric validation study could then determine whether the performance descriptors adequately capture common partially-articulated correct responses, and are interpreted consistently by different assessors.

7.4. Judicious Use of Students' Everyday Science Knowledge and Naïve Conceptions

By pairing the master teachers' characterizations of the item content with the instructional sensitivity effect sizes, we concluded that rubric performance levels that can be fulfilled with knowledge from everyday life experience tended to yield scores with negligible instructional sensitivity. While all the master teachers predicted that learners would utilize their daily life experiences to interpret the test items and formulate their responses (Hawkins & Pea, 1987; Zimmerman & Bell, 2014), they also noted that this cognitive pattern could be a barrier to responding with a precise explanation. The master science teachers gave examples that showed they could leverage students' everyday science knowledge to build deeper, more precise understanding.

Cultural fit of an assessment procedure—its coherence with regional communication patterns, experiences, values, or socioeconomic and cultural background—interacts with the epistemological beliefs that allow students to conceptualize and demonstrate their science knowledge (e.g., Klassen, 2006; Solano-Flores & Nelson-Barber, 2001). While certain cognitive patterns—such as reliance on familiar contexts to interpret scientific concepts—appeared across education systems, the specific manifestations of these patterns differed depending on students' prior environmental exposure, the curriculum structure, and the dominant instructional methods in their school. For instance, students who live in urban environments may have less familiarity with agricultural concepts than their rural counterparts do, affecting their ability to respond accurately to certain test items. This finding from the teachers' item content analysis underscores the importance of selecting item contexts that will be universally accessible to students from diverse backgrounds.

In three of the items' rubrics, the partially-correct score category assigned credit for a common naïve conception—of energy transfer, genetic heritability, or gravity. These score categories had negligible-to-moderate instructional sensitivity, but were too few to draw any general conclusion. It is possible that assessing students' naïve conceptions of science phenomena, whether they are treated as incorrect or partially correct responses, could inform education systems' curriculum decisions (Krajcik & Shin, 2023; Shepard, 2018).

Recommendations: Teachers across the national education systems generally favored including test items that are grounded in students' everyday experiences, such as direct observation of the natural world and common life events. However, because such items tend to exhibit similar difficulty levels regardless of prior instruction, test developers should carefully control the proportion of these items to maintain the intended meaning of the test scores. Cultural differences in what constitutes everyday knowledge and how students utilize it (Dawson, 2019) suggest that item instructional sensitivity should not be assumed to generalize across cultural contexts. Instead, future research within cultures could investigate if test items effectively measure curriculum-based learning rather than informal knowledge acquisition. Acknowledging these cultural variations can lead to refinements during item writing, particularly in large-scale international assessments like TIMSS, where item comparability across education systems is a key consideration.

7.5. Involve Master Educators in Test Development

Collaborating with experienced science educators during test development can help identify potential ambiguities in the instructions, item wording, and scoring rubrics. For instance, one item containing a labeled diagram was described by teachers from all four education systems as likely to confuse students. Other test items contexts were perceived as foreign to students, and might hinder their ability to transfer their knowledge effectively. One effective practice could be implementing structured teacher review panels that evaluate items for clarity, cultural appropriateness, and curriculum alignment. Such panels could identify ambiguous phrasing or visual representations that may not be evident during the initial item development. This approach aligns with research about collaborative test development, which suggests that educator involvement can improve item quality and reduce concealed sources of bias (Xerri & Briffa, 2018). Given the cross-national scope of TIMSS, particularly, incorporating educators from diverse education systems into the review process could potentially enhance the fairness of the test item content.

Ultimately, applying item instructional sensitivity guidelines during item writing, anchor item selection or form assembly could provide additional evidentiary support for a claim that the test scores are measuring students' curricular achievement. That is, implementing item IS principles or an index calculation would supply further validation evidence to support the intended meaning of the test scores. Recognizing that IEA explicitly cautions against causal interpretations of the TIMSS assessment results, evidence that students' scores reflect their learning in school may provide a stronger basis for participating education systems to generate hypotheses that they can test in their own regional context to inform education policymaking (Raudenbush & Kim, 2002). More broadly, integrating psychometric analyses with qualitative insights from educators can support the development of more equitable, culturally responsive, and instructionally sensitive assessments that better capture students' cognitive processes and educational experiences.

Acknowledgements

Dr. Gaelebala Nunu Tsheko and **Dr. Fiji Phuti-Kelebonye** contributed to data collection and thematic analysis, and **Kopo Neo Keaikitse** to data collection in Botswana. **Yueh-Chang Li** contributed to data collection and thematic analysis, and **Hung-Wei Liu** to data collection in Chinese Taipei. **Karen Alvarez Julian** contributed to thematic analysis in the United States.

References

Aikenhead, G. S., & Jegede, O. J. (1999). Cross-cultural science education: A cognitive explanation of a cultural phenomenon. *Journal of Research in Science Teaching*, *36*(3), 269–287.

Albano, A. D., & Rodriguez, M. C. (2018). Item development research and practice. In S. N. Elliott et al. (Eds.), *Handbook of accessible instruction and testing practices: Issues, innovations, and applications* (pp. 181–198). Springer.

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

Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education*, *95*(4), 639–669.

Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, *9*(1), 3–26.

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, *3*(22).

Brown, P. L., & Concannon, J. P. (2016). Students' perceptions of vocabulary knowledge and learning in a middle school science classroom. *International Journal of Science Education*, *38*(3), 391–408.

Burstein, L. (1989). *Conceptual considerations in instructionally sensitive assessments* (CSE Technical Report 333). Center for Research on Evaluation, Standards, and Student Testing. Los Angeles, CA: UCLA.

Cervetti, G. N., Hiebert, E. H., Pearson, P. D., & McClung, N. A. (2015). Factors that influence the difficulty of science words. *Journal of Literacy Research*, *47*(2), 153–185.

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., …, & Steele, C. M. (2013). *Criteria for high-quality assessment*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Dawson, E. (2019). *Equity, exclusion and everyday science learning: The experiences of minoritised groups*. London: Routledge.

Floden, R. E. (2002). The measurement of opportunity to learn. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 231–266). Washington, DC: National Academy Press.

Gómez Benito, J., Sireci, S. G., Padilla García, J. L., Hidalgo Montesinos, M. D., & Benítez Baena, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104–109.

Gough, B., & Madill, A. (2012). Subjectivity in psychological research: From problem to prospect. *Psychological Methods*, *17*(3), 374–384.

Hawkins, J., & Pea, R. D. (1987). Tools for bridging the cultures of everyday and scientific thinking. *Journal of Research in Science Teaching*, *24*(4), 291–307.

Ing, M. (2018). What about the “instruction” in instructional sensitivity? Raising a validity issue in research on instructional sensitivity. *Educational and Psychological Measurement*, *78*(4), 635–652.

International Test Commission. (2019). *ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations*. https://www.intestcom.org/files/guideline_diverse_populations.pdf

Klassen, S. (2006). Contextual assessment in science education: Background, issues, and policy. *Science Education*, *90*(5), 820–851.

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. Chicago, IL: University of Chicago Press.

Krajcik, J., & Shin, N. (2023). Student conceptions, conceptual change, and learning progressions. In N. G. Lederman, D. L. Zeidler, & J. S. Lederman (Eds.), *Handbook of research on science education* (Vol. III, pp. 121–157). Routledge.

Lakin, J. M. (2014). Test directions as a critical component of test design: Best practices and the impact of examinee characteristics. *Educational Assessment*, *19*(1), 17–34.

Leighton, J. P. (2019). The risk–return trade-off: Performance assessments and cognitive validation of inferences. *British Journal of Educational Psychology*, *89*(3), 441–455.

Levitt, H. M., Surace, F. I., Wu, M. B., Chapin, B., Hargrove, J. G., Herbitter, C., Lu, E. C., Maroney, M. R., & Hochman, A. L. (2022). The meaning of scientific objectivity and subjectivity: From the perspective of methodologists. *Psychological Methods*, *27*(4), 589–605.

Li, H., Qin, Q., & Lei, P. W. (2017). An examination of the instructional sensitivity of the TIMSS math items: A hierarchical differential item functioning approach. *Educational Assessment*, *22*(1), 1–17.

Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, *54*(8), 1024–1048.

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn*. Cambridge: Cambridge University Press.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement*, *36*(3), 217–232.

Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*(1), 1–22.

Naumann, A., Hochweber, J., & Klieme, E. (2016). A psychometric framework for the evaluation of instructional sensitivity. *Educational Assessment*, *21*(2), 89–101.

Naumann, A., Musow, S., & Katstaller, M. (2020). Instructional sensitivity as a prerequisite for determining the effectiveness of interventions in educational research. In H. Astleitner (Ed.), *Intervention research in educational practice: Alternative theoretical frameworks and application problems* (pp. 147–170). Waxmann.

Naumann, A., Rieser, S., Musow, S., Hochweber, J., & Hartig, J. (2019). Sensitivity of test items to teaching quality. *Learning and Instruction*, *60*, 41–53.

Nuthall, G., & Alton-Lee, A. (1995). Assessing classroom learning: How students use their knowledge and experience to answer classroom achievement test questions in science and social studies. *American Educational Research Journal*, *32*(1), 185–223.

Özdem-Yilmaz, Y., Cakiroglu, J., Ertepinar, H., & Erduran, S. (2017). The pedagogy of argumentation in science education: Science teachers' instructional practices. *International Journal of Science Education*, *39*(11), 1443–1464.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29*(4), 3–14.

Raudenbush, S. W., & Kim, J. S. (2002). Statistical issues in analysis of international comparisons of educational achievement. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 267–294). Washington, DC: National Academy Press.

Rodriguez, M. C. (2017, April). *Item and test design considering instructional sensitivity*. Paper presented at the National Council on Measurement in Education Annual Meeting, San Antonio, Texas.

Rottman, B. M., & Keil, F. C. (2011). What matters in scientific explanations: Effects of elaboration and content. *Cognition*, *121*(3), 324–337.

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics*, *24*(3), 293–322.

Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M. C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching*, *49*(6), 691–712.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369–393.

Samarapungavan, A. (2018). Construing scientific evidence: The role of disciplinary knowledge in reasoning with and about evidence in scientific practice. In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation* (pp. 56–76). Routledge.

Schmidt, W. H., & Maier, A. (2012). Opportunity to learn. In G. Sykes, B. Schneider, D. N. Plank (Eds.), *Handbook of education policy research* (pp. 541–559). Routledge.

Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied*

Measurement in Education, *31*(2), 165–174.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 553–573.

Terry, G., Hayfield, N., Clarke, V., & Braun, V. (2017). Thematic analysis. In C. Willig & W. S. Rogers (Eds.), The SAGE handbook of qualitative research in psychology (2nd ed.; pp. 17–37). London: Sage.

Traynor, A., Li, C.-H., & Zhou, S. (2025). Examining the instructional sensitivity of constructed-response achievement test item scores. *Educational and Psychological Measurement*.

Xerri, D., & Briffa, P. V. (Eds.). (2018). *Teacher involvement in high-stakes language testing*. Germany: Springer International.

Wang, C.-Y. (2015). Scaffolding middle school students' construction of scientific explanations: Comparing a cognitive versus a metacognitive evaluation approach. *International Journal of Science Education*, *37*(2), 237–271.

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design.* Oxford: Oxford University Press.

Zimmerman, H. T., & Bell, P. (2014). Where young people see science: Everyday activities connected to science. *International Journal of Science Education*, *4*(1), 25–53.

Zwick, R., Ye, L., & Isham, S. (2018). Aggregating polytomous DIF results over multiple test administrations. *Journal of Educational Measurement*, *55*(1), 132–151.

Table 1

Summary Statistics for Student Science Achievement and Teacher Instructional Questionnaire Items, by Country, 2011 and 2015

Country	Maximum Proportion of Students Tested in English	Minimum Number Tested in English	Median of Mean Item Scores†	Range in Proportions of Students in OTL Category for any Content Topic	Tested in 2011	Tested in 2015
Australia	1	6,343	.46	.23 – .97	X	X
Bahrain	.27	485	.45	.62 – .99	X	X
Botswana	1	2,818	.21	.26 – .99	X	X
Canada	.65	2,696	.50	.46 – .96		X
England	1	2,615	.52	.58 – .96	X	X
Ghana	1	5,725	.12	.49 – 1.00	X	
Ireland	.98	1,460	.50	.63 – .99		X
Kuwait	.07	140	.39	.50 – 1.00		X
Lebanon	.36	437	.25	.75 – 1.00	X	X
Malta	1	1,520	.37	.70 – .97		X
New Zealand	1	3,269	.47	.42 – .95	X	X
Oman	.19	789	.42	.31 – .97	X	X
Qatar	.47	927	.40	.69 – .99	X	X
Singapore	1	2,833	.66	.47 – .95	X	X
South Africa	.84	6,590	.16	.42 – .99	X	X
United Arab Emirates	.48	5,198	.42	.46 – .95	X	X
United States	1	6,564	.44	.56 – .97	X	X

Note. OTL = Opportunity to learn. † Median of mean item scores from 2015, when all 32 items were administered.

Table 2

Descriptive Statistics for 20 Released Constructed-Response TIMSS English-Version Grade 8 Science Test Items: 2011

Content Topic	Label	Num. of Score Points	Weighted Mean Score	Number of Test-taker Responses	Weighted Proportion of Test-takers in OTL Category
Cells and Their Functions	52263A	1	.19	5,186	.92
Ecosystems	42300A	1	.46	6,189	.78
Ecosystems	52303B	1	.40	6,570	.78
Reproduction and Heredity	42319	1	.38	5,270	.69
Reproduction and Heredity	52265	1	.38	6,435	.68
Acids and Bases	52043Z	1	.25	6,332	.85
Classification and Composition of Matter	52145	1	.31	6,227	.86
Chemical Change	42104	1	.21	6,348	.69
Mixtures and Solutions	42088	1	.47	6,391	.85
Physical and Chemical Properties of Matter	52049	2	.59	6,356	.86
Physical and Chemical Properties of Matter	42094	1	.37	6,000	.85
Physical and Chemical Properties of Matter	42228A	1	.20	5,098	.85
Physical and Chemical Properties of Matter	42228C	1	.42	5,624	.86
Electricity and Magnetism	42195	1	.10	5,442	.62
Energy Transformation and Transfer	42400	1	.20	5,645	.86
Energy Transformation and Transfer	42402	1	.19	6,552	.86
Forces and Motion	42293B	1	.63	5,837	.78
Forces and Motion	52141	1	.72	6,174	.77
Forces and Motion	52233	1	.16	6,256	.79
Light and Sound Energy	42273	1	.44	5,450	.75
Earth in the Solar System	52110	1	.35	6,130	.65
Earth's Processes, Cycles, and History	42301	1	.61	5,460	.59
Earth's Processes, Cycles, and History	42135	1	.44	5,948	.60
Earth's Structure and Physical Features	42406	1	.31	5,421	.59
Earth's Structure and Physical Features	52116	2	1.00	5,663	.58
Earth's Structure and Physical Features	52289C	1	.37	5,209	.58

Note. OTL = Opportunity to learn. Num. = Number. “Weighted” statistics were computed using the SENWGT sampling probability weight.

Table 3

Descriptive Statistics for 32 Released Constructed-Response TIMSS English-Version Grade 8 Science Test Items: 2015

Content Topic	Label	Num. of Score Points	Weighted Mean Score	Number of Test-taker Responses	Weighted Proportion of Test-takers in OTL Category
Cells and Their Functions	62098A	2	1.17	5,133	.96
Cells and Their Functions	62098B	2	.63	5,007	.96
Cells and Their Functions	62274	2	.81	5,148	.96
Cells and Their Functions	52263A	1	.24	4,701	.96
Ecosystems	42300A	1	.50	5,577	.83
Ecosystems	52303B	1	.42	5,699	.83
Life Cycles, Reproduction, and Heredity	42319	1	.36	4,857	.79
Life Cycles, Reproduction, and Heredity	52265	1	.38	5,716	.79
Acids and Bases	52043Z	1	.28	5,718	.95
Classification and Composition of Matter	52145	1	.39	5,482	.95
Chemical Change	42104	1	.21	5,629	.77
Mixtures and Solutions	42088	1	.50	5,636	.89
Physical and Chemical Properties of Matter	52049	2	.65	5,559	.95
Physical and Chemical Properties of Matter	42094	1	.38	5,389	.90
Physical and Chemical Properties of Matter	42228A	1	.23	4,567	.95
Physical and Chemical Properties of Matter	42228C	1	.52	4,908	.95
Electricity and Magnetism	42195	1	.16	4,792	.70
Electricity and Magnetism	62043	1	.33	5,289	.71
Energy Transformation and Transfer	42400	1	.29	5,128	.89
Energy Transformation and Transfer	42402	1	.20	5,689	.90
Forces and Motion	42293B	1	.17	5,682	.85
Forces and Motion	52141	1	.27	5,675	.85
Forces and Motion	52233	1	.16	5,477	.84
Light and Sound Energy	42273	1	.55	5,126	.75
Physical States and Changes in Matter	62005	1	.42	5,240	.86
Earth in the Solar System	52110	1	.31	5,505	.75
Earth’s Processes, Cycles, and History	42301	1	.62	5,140	.74
Earth's Processes, Cycles, and History	42135	1	.50	5,373	.74

Earth's Processes, Cycles, and History	62173A	1	.47	5,480	.74
Earth’s Resources, Their Use, and Conservation	62175	1	.45	4,704	.81
Earth’s Structure and Physical Features	42406	1	.31	5,252	.77
Earth's Structure and Physical Features	52116	2	1.09	5,186	.76
Earth's Structure and Physical Features	52289C	1	.43	4,794	.76

Note. OTL = Opportunity to learn. Num. = Number. “Weighted” statistics were computed using the SENWGT sampling probability weight.

Table 4

Item Difficulty Difference Instructional Sensitivity Index Values for 32 Constructed-Response TIMSS English-version Grade 8 Science Assessment Items, each Administered during Two Years: 2011 to 2019

Content Topic	Label	Num. of Score Cat.	Item Discrim. Coefficient 2015	Step 0 → 1 BDIFF 2011	Step 1 → 2 BDIFF 2011	Step 0 → 1 BDIFF 2015	Step 1 → 2 BDIFF 2015	Step 0 → 1 BDIFF 2019	Step 1 → 2 BDIFF 2019
Biology									
Cells and Their Functions	62098A	3	1.52			.07 [-.28, .42]	.11 [-.19, .40]	.15 [-.38, .68]	.10 [-.51, .71]
Cells and Their Functions	62098B	3	1.45			**.52** [.13, .90]	-.12 [-.55, .31]	.34 [-.05, .71]	.05 [-.31, .41]
Cells and Their Functions	62274	3	.82			-.06 [-.31, .19]	.15 [-.22, .52]	-.11 [-.99, .77]	.17 [-.90, 1.24]
Cells and Their Functions	52263A	2	1.46	.07 [-.16, .29]		.24 [-.16, .64]			
Ecosystems	42300A	2	1.72	-.06 [-.19, .06]		.01 [-.16, .17]			
Ecosystems	52303B	2	1.20	-.19 [-.33, -.04]		-.03 [-.20, .14]			
Life Cycles, Reproduction and Heredity	42319	2	2.21	-.10 [-.22, .04]		-.09 [-.21, .03]			
	52265	2	1.66	.11		.04			

Life Cycles, Reproduction and Heredity				[-.01, .23]		[-.10, .18]	
Chemistry							
Physical and Chemical Properties of Matter	52049	3	1.21	**.31**	**.49**	**.43**	**.65**
				[.13, .50]	[.24, .74]	[.10, .76]	[.17, 1.14]
Acids and Bases	52043Z	2	.86	**.64**		**.83**	
				[.47, .86]		[.45, 1.21]	
Chemical Change	42104	2	1.30	.14		**.28**	
				[-.01, .27]		[.10, .46]	
Classification and Composition of Matter	52145	2	1.83	**.45**		**.42**	
				[.32, .58]		[.24, .61]	
Mixtures and Solutions	42088	2	.79	.09		**.36**	
				[-.12, .29]		[.05, .67]	
Physical and Chemical Properties of Matter	42094	2	1.05	**.38**		**.49**	
				[.20, .56]		[.22, .75]	
Physical and Chemical Properties of Matter	42228A	2	2.07	**.69**		**.39**	
				[.49, .88]		[.09, .68]	
Physical and Chemical Properties of Matter	42228C	2	1.95	**.41**		.12	
				[.26, .55]		[-.12, .37]	
Earth Science							
Earth's Structure and Physical Features	52116	3	1.32	-.02	-.05	.06	.01
				[-.16, -.12]	[-.17, .08]	[-.11, .24]	[-.16, .17]

Earth in the Solar System	52110	2	1.55	**.35** [.22, .49]	**.36** [.20, .52]	
Earth's Processes, Cycles, and History	42135	2	1.37	**.22** [.10, .34]	**.20** [.06, .34]	
Earth's Processes, Cycles, and History	42301	2	1.58	**.24** [.11, .38]	**.19** [.03, .34]	
Earth's Processes, Cycles, and History	62173A	2	.82		.03 [-.17, .22]	.20 [-.03, .44]
Earth's Resources, Their Use, and Conservation	62175	2	.88		-.04 [-.31, .22]	-.02 [-.30, .25]
Earth's Structure and Physical Features	42406	2	1.69	-.04 [-.16, .08]	-.01 [-.16, .15]	
Earth's Structure and Physical Features	52289C	2	1.00	**.17** [.02, .33]	.03 [-.14, .21]	
Physics						
Electricity and Magnetism	42195	2	.77	**1.25** [1.01, 1.48]	**.71** [.41, 1.02]	
Energy Transformation and Transfer	42400	2	1.49	**.46** [.28, .63]	**.70** [.43, .98]	
Energy Transformation and Transfer	42402	2	1.37	.10 [-.11, .30]	**.35** [.09, .61]	

Forces and Motion	42293B	2	1.26	.11 [-.05, .27]	**.29** [.05, .53]	
Forces and Motion	52141	2	2.14	.08 [-.05, .21]	**.25** [.09, .42]	
Forces and Motion	52233	2	.92	-.05 [-.24, .13]	.17 [-.07, .42]	
Light and Sound Energy	42273	2	1.80	**.42** [.28, .56]	**.50** [.37, .64]	
Physical States and Changes in Matter	62005	2	1.49		**.33** [.09, .57]	**.34** [.04, .63]

Note. OTL = Opportunity to learn; BDIFF = Difference in category difficulty parameters between "No OTL" and "OTL" groups; Discrim. = Item discrimination coefficients are reported from a single-group model for the 2015 item response data. Effect sizes printed in boldface are significantly different from 0.

Table 5

Item Difficulty Difference Instructional Sensitivity Index Values for 32 Constructed-Response TIMSS 2011 Grade 8 Science Assessment Items, by Test Language Source Version

Content Topic	Label	Number of Categories	BDIFF Arabic Version	BDIFF English Version
Adaptation and Variation in Survival of Species	32451	3	.01	.05
Major Organs and Organ Systems of Organisms	32306	3	-.25	**-.44**
Life Cycles, Reproduction, and Heredity	32530Z	3	.29	.15
Life Cycles, Reproduction, and Heredity	42297	3	-.01	**-.26**
Physical and Chemical Properties of Matter	52049	3	**.33**	**.43**
Classification and Composition of Matter	42305	3	**.91**	**.69**
Chemical Change	42100	3	**.25**	.15
Physical States and Changes in Matter	42173Z	3	.18	.06
Physical States and Changes in Matter	42404	3	.13	.01
Light and Sound Energy	32369	3	**.36**	**.38**
Earth's Structure and Physical Features	32650Z	3	-.03	-.08
Earth's Processes, Cycles, and History	42317	3	.21	.15
Earth's Structure and Physical Features	52116	3	**.28**	-.03
Life Cycles, Reproduction, and Heredity	42319	2	-.02	-.10
Life Cycles, Reproduction, and Heredity	52265	2	.25	.09
Cells and Their Functions	52263A	2	.10	.07
Ecosystems	42300A	2	-.17	-.06
Acids and Bases	52043Z	2	.16	**.64**
Classification and Composition of Matter	52145	2	.15	**.45**
Mixtures and Solutions	42088	2	-.02	.09
Chemical Change	32679	2	**.55**	.16
Chemical Change	42104	2	.10	.14
Physical and Chemical Properties of Matter	42094	2	**.62**	**.38**
Energy Transformation and Transfer	42400	2	.19	**.46**
Light and Sound	42273	2	**.27**	**.40**

Electricity and Magnetism	42195	2	.02	**1.24**
Forces and Motion	42293B	2	.11	.12
Forces and Motion	52141	2	-.06	.06
Earth's Structure and Physical Features	42406	2	.08	-.04
Earth's Structure and Physical Features	52289C	2	.05	.18
Earth's Processes, Cycles, and History	42301	2	.05	**.27**
Earth in the Solar System	52110	2	.23	**.35**

Note. OTL = Opportunity to learn; BDIFF = Difference in item difficulty parameters between "No OTL" and "OTL" groups. Effect sizes printed in boldface are significantly different from 0.

Table 6

List of TIMSS Science Achievement Test Items Selected for Teachers' Content Analysis, Sorted by Instructional Sensitivity Index Magnitude

Instructional Sensitivity Index = High	Instructional Sensitivity Index = Negligible
Biology	
32530Z: Reproduction and Heredity	32451: Role of Variation and Adaptation in Survival of Species
52092Z: Differences among Major Taxonomic Groups	52265: Life Cycles, Reproduction, and Heredity
	62274: Ecosystems
Chemistry	
42305: Classification and Composition of Matter	42088: Mixtures and Solutions
52043Z: Acids and Bases	42100: Chemical Change
52145: Classification and Composition of Matter	
Earth Science	
42301: Earth's Processes, Cycles, and History	42406: Earth's Structure and Physical Features
52110: Earth in the Solar System	52116: Earth's Structure and Physical Features
	62234: Earth's Resources, Their Use, and Conservation
Physics	
42094: Physical States and Changes in Matter	42404: Physical States and Changes in Matter
42195: Electricity and Magnetism	52233: Forces and Motion
42273: Light and Sound Energy	

Figure 1
Classification of matter item (S042305) with high score category *BDIFF* = .85, .37

The table below shows some elements, compounds, and mixtures.

Classify them by putting an X in the appropriate column beside each one.

	Element	Compound	Mixture
Air			
Sugar			
Salt			
Gold			
Sea Water			
Helium			

Table 7

Master Science Teachers' Demographic Information

Education System	Participant ID	Years of Science Teaching Experience	School Type
Botswana	T1 T2	16 26	Public schools
Chinese Taipei	T3 T4	22 35	Public schools in the Southern region
Türkiye	T5 T6	13 9	Public schools
United States	T7 T8	10 11	Private and public schools in one State

Table 8

Summary of Test Item and Rubric Design Features that Affect Item Score Instructional Sensitivity

Instructional sensitivity generally **varies for score categories** within an item. The rubric's performance level descriptions partly explain these differences.

1) Use of **scientific vocabulary** in the item text or a rubric performance level (i.e., score category) description seems to increase the IS of item scores.

2) Use of an **academic science concept** in the item, together with a rubric performance level description that requires explanation or **elaboration** of responses is associated with high IS of item scores.

3) Rubric performance levels that can be fulfilled with knowledge from **everyday life experience**, shared by many students in many countries, appear to yield item scores with negligible IS.

4) When a specific misinterpretation of an item's visual material (e.g., diagram, graph) or text that is unrelated to science achievement is highly probable (**lack of clarity**), the item scores tend to have negligible IS.

5) Rubric performance level descriptions that evoke common **naïve conceptions** of science phenomena do not tend to produce item scores with high IS (but may be diagnostically useful).

Appendix 1

Semi-Structured Item Content Analysis Interview Protocol Example (US English Version)

Starting Instructions: To evaluate the content quality of the TIMSS science test items, I encourage you to consider all your experiences teaching science to Grade 8 students at this school and other schools where you have taught (if any). Since students can interact with one test item in different ways, please consider many students who you have taught.

*[**Interviewer Note**: We will show test item 1 to the participant, ask questions and record their responses, show the rubric for test item 1, ask questions and record their responses, then show test item 2, and continues this pattern.]*

[Present Item 1…]

1) What do you expect students will understand as the meaning of this item? How will students interpret the item's question?

2) How will students think about their answer for this item?

3) What types of answers do you predict students will give?

During the next part, I will ask you to think about students who have and have not had previous formal science instruction (at school) about the item's "broad content topic." If a student has been taught about, for example, Chemical Change or Reproduction of Life, we consider them to have had instruction about the item's content topic. To give examples about the breadth of one "topic," here is a list [Question 20 from the TIMSS teacher survey questionnaire]. We are focused on regular science instruction following the local curriculum. We are not interested in situations when educators directly teach how to solve a specific test item. We are interested in how students who have not been taught will interact with each item.

[Present rubric for Item 1…]

Considering the item, and score categories in the rubric, please make a prediction:

4) How do you anticipate students will use content that they have learned at school to answer the question? [Follow up: Are there other ways students may use content that they have learned at school to answer?] [Follow up for 2-point items: How might students use content that they have learned at school to develop a partially correct answer?]

5) If a student has not learned about this idea at school, are there ways that that they could correctly answer this item? How?

6) Among all students with no previous formal science instruction about the item's broad content topic, how probable is it that they will write answers in the (fully)

High probability	Moderate probability	Low probability	Very low probability

correct category?

7) Are there ways that students without instruction can develop a partially correct answer?

8) Among all students with no previous formal science instruction about the item's broad content topic, how probable is it that they will write answers in the *partially correct* category?

High probability	Moderate probability	Low probability	Very low probability

[Present Item 2...]

Item Comparison Questions (after last item):

Consider Items 52265, 42100, 52116, and 522336 as Item Set A and Items 52092Z, 42305, 52110, and 42195 as Item Set B:

- Are there any similarities in the types of thinking required for students to correctly answer the items in Item Set 1A? In Item Set 1B?
- Comparing the two item sets, are there any differences in the types of thinking required to correctly answer them?