Analysis of Student and Item Performance on Three-Dimensional Constructed Response Assessment Tasks

Brian D. Gane, University of Illinois at Chicago
Kevin W. McElhaney, SRI International
Sania Z. Zaidi, University of Illinois at Chicago
James Pellegrino, University of Illinois at Chicago

Paper presented at the 2018 NARST Annual International Conference, Atlanta, GA

Part of a related paper set: *Designing and Implementing Instructionally Supportive Assessment Tasks for Promoting Three-Dimensional Learning for all Students*

Suggested citation:

Gane, B. D., McElhaney, K.W., Zaidi, S. Z., Pellegrino, J. W. (2018, March). *Analysis of student and item performance on three-dimensional constructed response assessment tasks*. Paper presented at the NARST Annual International Conference, Atlanta, GA.

To address the vision of the NGSS, instructionally supportive assessments are needed that can foster students' integrated, three-dimensional learning (Pellegrino, Wilson, Koenig, & Beatty, 2014). The focus of assessments, and the nature of the knowledge, skills, and abilities that they target, inevitably become the targets of instruction and the focus of what teachers teach and what students learn. Crucially, given the major shifts in the science education landscape brought about by the NGSS, well-designed assessments can act as clear signals to teachers and students about what they are expected to learn (Pellegrino, 2013). Assessment of multi-dimensional thinking is challenging to implement in traditional multiple-choice formats due to the limited nature of the student response (Gorin & Mislevy, 2013). Even when they are carefully designed and linked to students' misconceptions or learning trajectories (e.g., Briggs, Alonzo, Schwab, & Wilson, 2006; Herrmann-Abell, Flanagan, & Roseman, 2013; Sadler, 1998) multiple-choice formats are still difficult to use as evidence of students' *knowledge-in-use* (i.e., their ability to use their content knowledge to engage in NGSS practices, such as developing explanations or models). Constructed response item formats, on the other hand, allow students to demonstrate various aspects of proficiency, including their ability to integrate the three NGSS dimensions. Further, such constructed response formats support the use of multiple scores and/or partial credit. However, developing valid, instructionally useful constructed response assessments requires considering multiple aspects of validity, including whether student response variation can be modeled statistically (Pellegrino, DiBello, & Goldman, 2016). In this paper we report preliminary results from an assessment validation study with over 800 middle school students. We focus on understanding student and task performance on tasks aligned to aspects of two NGSS performance expectations. Our research question is *Can model-based information about student and item[1] performance be provided as warrants for the inferential validity of the assessment tasks, individually and collectively?*

Background: Validating Assessments

Assessment is a multi-faceted process in which multiple design and practical elements need to be considered concurrently. One model for reasoning about assessments is the assessment triangle (Pellegrino, Chudowksy, & Glaser, 2001), whose vertices represent different aspects of an assessment. The assessment triangle has three vertices: cognition, observation, and interpretation. In this paper, we focus primarily on the interpretation vertex, which is how individuals (e.g., practitioners, researchers, etc.) make sense of those observations that arise from students' work on assessment tasks. Assessment is an inferential process, where one attempts to infer students' general competencies from their performance on a specific, limited set of tasks (Pellegrino et al., 2001). Further, assessments can be designed and used to yield different information depending on the type of inference one intends to make. Validating assessments therefore is not a process that can be applied uniformly to all types of assessments (nor is it an all-or-none statement); validation needs to be tailored to the type of inferences one is attempting to make. We view validation as an evidence-based argument (Kane, 2006; 2013) in which one makes claims about the use of assessment and these claims are backed by evidence. This evidence can be varied and arise from different sources. Some of this evidence may be provided by details of the design process if it is made transparent (e.g., McElhaney, Zaidi, Gane, Krajcik, Alozie, & Harris, 2018), rather than being designed inside a "black box". Additionally, empirical evidence needs to be collected that provides support for the reasonableness and appropriateness of the intended inferences.

[1] In this paper we use "item" and "task" interchangeably.

Pellegrino and colleagues (2016) provide a framework for validating assessments that are intended to support instruction. These are assessments that may be used by teachers in their classes to determine where students are with respect to different learning objectives, and to modify their instruction to ensure that students reach those learning objectives. Importantly, for these types of assessments teachers need to know what their students' strengths and weaknesses are with respect to the different learning objectives. Assessment tasks therefore need to be able to provide teachers with information that can be used to diagnose students' ability to productively engage with the content and to diagnose where students are struggling with problematic ideas or practices.

In this paper we report results from preliminary analysis from a validation study using the NGSA assessment tasks (https://ngss-assessment.portal.concord.org/) with middle school students (6th and 7th grade students). We are interested in showing evidence that the tasks can be used to reveal both individual features of student responses and as an overall measure of students' three-dimensional learning with regard to different NGSS performance expectations (PEs). To do this, we model student performance on the basis of student and item features and show how analysis can be used to reveal overall, aggregate performance on using disciplinary core ideas with a variety of practices and crosscutting concepts. Further, we show how analysis can be conducted within learning performances or items to show specific aspects of students' problematic or productive thinking.

Design & Procedure

Participants

This study was conducted during Spring 2016 in three school districts serving diverse student populations in Illinois, Oklahoma, and Wisconsin. District demographics for the 2015-2016 academic year are presented in Table 1. Seventeen 6th and 7th grade teachers and their students from 12 schools across the three districts participated. All teachers were using a commercially available NGSS-aligned curriculum in their instruction, *Investigating and Questioning our World through Science and Technology* (IQWST). IQWST focuses students on using scientific practices to explore and understand observed phenomenon (https://activatelearning.com/iqwst). Its emphasis on integrating multiple NGSS dimensions, especially with regard to the disciplinary core ideas and science practices, aligns it with the *Framework* (NRC, 2012) and aspects of the NGSS. Students had not previously used the NGSA task portal nor had they completed any of the assessment tasks prior to their participation.

Task Domain

Participants completed a variety of middle school physical science tasks that were associated with the chemical reaction topic. This topic includes two primary performance expectations: MS-PS1-2 (*Analyze and interpret data on the properties of substances before and after the substances interact to determine if a chemical reaction has occurred*) and MS-PS1-5 (*Develop and use a model to describe how the total number of atoms does not change in a chemical reaction and thus mass is conserved*) and a supporting performance expectation: MS-PS1-1

(*Develop models to describe the atomic composition of simple molecules and extended structures*).

During the NGSA design process, we develop learning performances that are derived from the target performance expectation(s) (Harris, Krajcik, Pellegrino, & McElhaney, 2016; McElhaney et al., 2018). These learning performances are smaller in scope than the performance expectation, often including only a portion of the disciplinary core idea (DCI) elements that are in the PE. These learning performances are three-dimensional, integrating a science practice and a crosscutting concept with the disciplinary core idea. However, the learning performances often differ from the PE in terms of the science practice and/or crosscutting concept used. This difference in science practices and crosscutting concepts allows teachers to assess students' proficiency with a variety of science practices and crosscutting concepts, consistent with the vision of the NGSS and *Framework*. The set of learning performances for these three target PEs are displayed in Table 2. The assessment tasks that were analyzed were aligned to one of these learning performances. By using a variety of assessment tasks from the set of learning performances we can analyze how students are able to integrate the three NGSS dimensions while they reason about and use their knowledge of chemical reactions (including characteristic properties of substances, rearrangement of molecules, generation of new substances, and conservation of mass).

Design

We used a matrix sampling design to collect data on the middle school physical science tasks that were associated with the chemical reaction topic. In total, there are 33 tasks associated with the chemical reaction topic that includes two primary performance expectations (MS-PS1-2; PS1-5) and one supporting performance expectation (MS-PS1-1). In this paper, we analyzed performance on a subset of 31 tasks[2].

Students were pseudo-randomly assigned 1 of 16 booklets; each booklet contained six assessment tasks. Students had one class period (approximately 60 minutes) to complete the assessment tasks. Students were instructed to start with the first task in their booklet, however students could skip tasks. Not all students completed all assigned tasks. Students completed tasks individually on a laptop, desktop, or tablet.

Scoring

Our scoring approach followed directly from our assessment task and rubric design framework, which is based on the focal knowledge, skills, and abilities (FKSAs) that are hypothesized to underlie task performance (Harris et al., 2016). Each learning performance has a set of FKSAs that were defined prior to developing assessment tasks and rubrics. Each learning performance has a different combinations of FKSAs. The same FKSA can appear in multiple learning performances, but each learning performance has a unique combination of FKSAs. Tables 3 and 4 present the FKSAs for two learning performances, LPC04 and C06. Note that FKSA J appears in both C04 and C06, demonstrating how FKSAs can be part of multiple learning performances. Tasks are developed to align to the learning performance and to elicit evidence of the FKSAs. Therefore, each task has all FKSA's of its parent learning performance (e.g., any task in learning

[2] One task (Task 20) had been dropped prior to analysis because of problems with the task design that became apparent during data scoring. One task (Task 35) was dropped prior to IRT analysis.

performance C04 will be designed to elicit evidence of FKSA G, H, I, and J. These details of the FKSA-learning performance-task are crucial for understanding our scoring model. Each task has a rubric that has multiple rubric components, one component for each FKSA. A rubric component is an element of the rubric that can be scored for its presence or absence in a student response. Each rubric component has an associated evidence statement (McElhaney et al., 2018).

Before applying the rubrics to student responses, raters went through a training process in which they learned how to use the rubrics and scored a sample of student responses to check initial agreement. Scoring a task involves scoring each of the distinct rubric components (i.e., FKSAs). Agreement between raters is checked by comparing the various rubric component scores in each task. After they completed this training and an initial agreement check, two raters independently scored subsets of student responses until they reached high agreement. After achieving agreement, one rater scored the remaining responses. Disagreements between raters were resolved through discussion.

Analysis & Findings

Agreement

We calculated inter-rater agreement to evaluate scoring of the FKSAs across multiple tasks within a learning performance. Overall agreement between raters was acceptable, differing by FKSA and learning performance (median Cohen's $K = .77$, range = .52, .93; median inter-rater agreement = 93%, range = 77%, 98%). The Cohen's K values need to be cautiously interpreted because of the unequal distribution of the rubric component scores (i.e., positive skew).

Tasks Attempted

Students were assigned booklets with six tasks, however, they did not necessarily complete all tasks within a booklet. Figure 1 shows the number of task attempts across the full sample of 887 students. We removed data from any students who completed less than two tasks[3]. We dropped these subjects due to problems that might occur during model estimation given the amount of missing data that would be present for these students. In total, 57 students were dropped, leaving a sample size of 830. We used this sample of 830 students for the remaining analyses reported in this paper.

The number of students that were assigned to specific booklets was unequal, with some booklets being assigned more or less frequently[4]. Further, students could skip tasks. These two factors caused an unequal number of student responses per task. Table 5 provides the number of responses for each of the 31 tasks and Figure 2 displays these data in a histogram. The majority of the tasks had > 100 responses per task. This is an adequate sample size for the types of psychometric analyses that we conducted.

Total Score

[3] Subjects that did not attempt the assessment task at all (no response to any of the task response prompts) were scored as missing data/non-attempts. Subjects that gave any attempt (even to respond such as "idk") were scored (although almost invariably these scores ended up as all 0's on the rubric components).

[4] This is primarily due to differences in the number of students that were in classes. Within each class three booklets were used. Students were randomly assigned one of these three booklets.

A task's total score was computed as the sum of the individual (dichotomous) scores on each task's rubric components. The task's score range is determined by the total number of rubric components present in the task[5]. Our set of tasks have between 2–4 FKSAs, therefore possible task total scores range from 0–4, depending on the learning performance. Tables 6 and 7 show two example student responses for the same task. The responses differ in terms of the FKSAs present in their answer, and therefore their total score differs also. Total score does not give an indication of *which* FKSA/rubric components were evidenced by a student's response. These data are available only by analyzing the components of total score–the individual FKSA variables. For example, learning performance C06 has three FKSAs. If a student gets a total score of 2 it might be because the student got credit for FKSAs J & L, or FKSAs J & M, or FKSAs L and M. Because FKSAs differ across learning performances, this fine-grained analysis is best performed individually for each learning performance.

IRT Analysis

We used Item Response Theory (IRT) (Embretson & Reisse, 2000) to analyze students' performance on the assessment tasks. The generalized partial credit model (GPCM) is appropriate for polytomous, ordinal data such as our assessment tasks' total score (Muraki, 1992). We used the R ltm package (Rizopoulos, 2006) to fit GPCM models to the data. The GPCM includes a slope parameter (α_i) for each item and category intersection parameters (β_{ij}) that indicate where on the latent trait continuum an individual is likely to transition from one category to another.

Early attempts at fitting models to the data indicated unrealistic item parameter estimates for Task 35 (e.g., a category intersection parameter > 700). We subsequently dropped Task 35 from the remaining analysis. In order to test model fit, we attempted to fit three models: a GPCM model with a Rasch constraint (all item slope parameters are set to 1), a GPCM model with a 1PL constraint (all item slope parameters are set to the same value, but this value is estimated rather than set to 1), and a GPCM model with a 2PL constraint (a different slope parameter for each item)[6]. The GPCM model with a 2PL constraint did not converge, and thus we did not analyze it further. The results for the two model fits are presented in Table 8. The p-value in Table 8 is the result of an ANOVA test on the Rasch vs. 1PL model: There was no significant difference in model fit between the two models. Given these results we selected the Rasch model for further analysis. We performed a goodness of fit test using parametric bootstrap estimation[7] and the resulting p-value for the comparison between observed (χ_{obs} = 23287.24) and simulated chi-square values was non-significant, $p = .17$. This provides support for fitting the GPCM model with a Rasch constraint to the data. More generally, this shows that students' performance on an item (i.e., total score) can be modeled as a function of student parameters (θ) and item parameters (α and β).

We analyzed test information and found that the information value was highest for higher ability students, peaking near theta = +1.5 (see Figure 3). Additionally, the information value was

[5] Many rubric components are scored as (0, 1) but ooccasionally rubric components have two levels (0, 1, 2). In this analysis we converted polytomous FKSA scores to be dichotomous (e.g., 2 → 1).

[6] These first two models are equivalent to the Partial Credit Model whereas the third model is the full Generalized Partial Credit Model.

[7] 100 bootstrap samples were computed

greater for positive theta values than for negative theta values: information in (-6 to 0) = 18.29 (19.7%); information in (0 to 6) = 73.9 (79.4%). This shows that overall, this set of items is most precise at estimating the ability of the higher ability students in our sample. In general, we found that the items were difficult for our sample of students. This makes sense because many of the students in our sample were in their first year of taking a science course that was (beginning to be) aligned to the NGSS. We can expect that students that have gone through multiple years of study with science courses aligned to the NGSS would perform better. Our data show that there is room for students' proficiency estimates to increase, and as they increase we will see students better prepared to complete our tasks. Thus, although our tasks are difficult for the current sample, and the information curve peaks to the right of $\theta = 0$, we expect that follow-up studies would show the information curve peak moving to the left.

A Closer Look at Two Learning Performances

For the remainder of this paper we focus on tasks from two learning performances[8]: LPC04 (*evaluate* whether a model explains that a chemical reaction produces new substances and conserves atoms) and LPC06 (*develop* a model of a chemical reaction that explains new substances are formed by the regrouping of atoms, and that mass is conserved). These learning performances are interesting because they both involve students working with models of a chemical reaction, reasoning about conservation of matter. In LPC04 students are evaluating existing models, whereas in LPC06 students are creating the models themselves. During task development we hypothesized that the it might be more difficult for students to engage in creating a model than it would be to evaluate a model. LPC04 has three FKSAs (see Table 3) and LPC06 has four FKSAs (see Table 4); one FKSA (FKSA J) is present in both learning performances.

First we examine the item information plots (see Figure 4). The information peaks between +1.5 and +2.0 for these items. There is no apparent pattern of difference between the curves for LPC04 and those for LPC06. The items in both learning performances have higher discrimination (lower measurement error) for high ability students in our sample.

Next we examined the item category response characteristic curves (see Figures 5 and 6 for items in LPC04 and LPC06, respectively). Learning performance C04 has four rubric components, therefore it has a total score range 0–4. Learning performance C06 has only three rubric components, therefore it has a total score range of 0–3. These curves indicate the probability (y-axis) that a student will have a specific total score (e.g., $x = 1$) based on the students' standing on the latent trait (θ; x-axis). In general, these data show a desirable item response pattern–increasing item score values align with differential and increasing estimates of overall student proficiency.

To start, we overview the results of the four tasks in LPC04 by surveying the item category response characteristic curves (Figure 5). When examining the item characteristic curves it is informative to identify where two curves intersect; the location of this intersection shows the difficulty associated with moving from one curve to the next (i.e., its "step difficulty"). For instance, in Task 26 the first category intersection occurs at approximately $\theta = 1$. The four items

[8] See McElhaney et al. (2018) for a description of how learning performances are derived from the NGSS performance expectations.

differ in where along the latent trait a category transition occurred (e.g., x = 1 → x = 2), indicating differences between tasks in terms of their overall difficulties. Further, one can examine whether each curve has a peak that is higher than all other curve peaks – this indicates it is probable to get student responses in this category. For tasks 27 and 28 there is a latent trait (θ) range that was associated with a probability of responding either 1, 2, or 3. For tasks 26 and 36 there was one score value (x = 3 in Task 26 and x = 2 in Task 36) that was not probable across the entire range of the latent trait. This pattern indicates that in LPC04 there was a difference among tasks in terms of the probability of students getting partial credit scores (e.g., x = 1–3).

Next we survey the item category response characteristic curves for LPC06. Whereas LPC04 had four FKSAs, LPC06 has three FKSAs, limiting the possible total score to three. In LPC06 we see a similar trend as in LPC04 with respect to differences in where the categories intersect across the different tasks. For instance, in Task 21 and 22 the first category intersection parameter is near $\theta = 0$, whereas for Task 25 and Task 32 the first category intersection parameter is between $\theta = +1.5$ to +2.5. These data indicate that for students with lower ability, they are more likely to get partial credit (1 of 3 FKSAs correct) in Tasks 21 and 22, relative to Task 25 and 32. A further examination of the tasks, including the scaffolding present in the tasks and the response demands, is warranted to better understand this difference between tasks in LPC06. Nevertheless, these data show important information about the tasks that can be used to inform their use in classrooms. Teachers using these data might decide to use Tasks 21 or 22 early in instruction when students are still learning, and use Task 25 or 32 later in instruction when students are more adept at creating models of chemical reactions that conserve mass.

There are two main patterns that emerge from the item characteristic curve data. First, item difficulties vary across items, with the category intersections having different theta locations. For example, Task 26 has its first category intersection at $\theta = 0$ whereas the equivalent category intersection for Task 36 is closer to $\theta = 1.5$. Items that have category intersections further to the right (i.e., a higher theta value) are more difficult. Interestingly, because this is a partial credit IRT model, the different partial credit scores have different difficulties associated with them; these can be seen on the individual item characteristic curve where two curves intersect (i.e., two different total score categories intersect). The second pattern which emerges is that some item characteristic curves have functions where the probability is less than all others (e.g., Task 36, x = 2 curve). In the GPCM the ordering of category intersections ("step difficulties") is not guaranteed; it is possible to have category intersections that are reversed. In fact, this reversal occurs for some of the items in our sample (e.g., Task 26, LPC04, Figure 5; Task 22, LPC06, Figure 6). Such a reversal indicates that it is harder to transition in lower categories then it is to transition in higher categories. Task 22 category threshold parameters ($\beta_{i1} = -0.04$, $\beta_{i2} = 1.90$, $\beta_{i3} = 1.36$) indicate that it is more difficult to move from a score of 1 to 2 ($\beta_{i2} = 1.90$) than it is to move from a score of 2 to 3 ($\beta_{i3} = 1.36$). Such a pattern might result from a difference in related FKSAs in an item: once a student gets 1 of 3 FKSAs correct it is not easy to get one of the other FKSAs correct. However, once a student can get a 2nd FKSA correct then it is easier to also get that 3rd FKSA correct. An examination of the response pattern of the FKSAs might reveal whether one FKSA was easier than the other two FKSAs, as suggested by this category intersection reversal. We conduct such an examination later in this paper.

To further understand these data, and to demonstrate the utility of our scoring approach, we analyzed one item (Task 26, LPC04; see Figure 7) in more depth. Task 26 has an item response pattern where there is a wide range of theta values where it is probable that a student will score a 1 or 2 and no range of theta values where it is probable that a student will earn a score of 3 (i.e., a category intersection reversal). The raw response data also bear out this pattern: few students in our sample earned a score of 3 or 4. Those that did earn a score of 3 or 4 likely have high latent trait estimates.

To examine Task 26 in further detail we analyzed the pattern of FKSAs that underlie the total score. As explained earlier, the item characteristic curves are calculated using total score that indicates the number of FKSAs that were present in a students' response, but not which FKSAs were present. Figure 8 plots the frequency of FKSA scores that were equal to one. FKSA H has the highest frequency, indicating it might be the "easiest" FKSA for students. FKSA J, on the other hand, has the lowest frequency, indicating it might be the "hardest" FKSA for students. Comparing this pattern to the content of the FKSAs (Table 3) yields insights. FKSA H and I require the student to make a determination of whether the presented chemical reaction model is adequate. FKSAs G and J require students to justify their response by describing why the model is adequate. FKSA J, specifically, asks students to do this with respect to whether the model shows conservation of mass. To do this, students have to state that the number of both nitrogen and hydrogen atoms before and after the chemical reaction are the same (alternately, students can state the number of atoms as long as they make clear that number is the same on each side of the chemical reaction, i.e., before and after the reaction). These data suggest that students are better able to evaluate a model showing that a chemical reaction produces new substances than they are at evaluating whether the model shows that mass is conserved.

Again using Task 26, we analyzed the correlations among FKSAs to see if that yielded additional insights (see Table 9). FKSAs H and I are correlated, both of these require students to determine whether the model is adequate. FKSA G and J are correlated, both of these require students to describe why/how the model adequately represents the chemical reaction. Finally, there is a significant correlation between whether students determine the model shows conservation of atoms (FKSA I) and whether they describe how the model represents this conservation of mass (FKSA J). This correlation is intuitive as well–if a student does not determine that the model correctly shows that atoms are conserved then the student will not likely explain how it shows that mass is conserved.

Although these analyses are for one task only, the data give insight into students' cognition. Students in our sample were less likely to be able to describe why a model was adequate than they were to decide whether it was adequate. Additionally, students in our sample tended not to describe that the model showed conservation of mass. These patterns are interesting but similar analysis needs to be completed on the other items in the learning performance. Nevertheless, these findings from one item have clear instructional uses: Teachers can use the assessment task to see whether students are having difficulty understanding that the model is adequate (FKSAs H & I) or whether they are having difficulty describing why the model is adequate (FKSAs G & J). Further, teachers can use the pattern of results (as shown in Figure 8) to understand whether their students are struggling with modeling that chemical reactions create new substances (FKSAs G & H) or that chemical reactions must conserve mass (FKSAs I & J).

Discussion

This paper begins to sketch a framework for validating our assessment tasks with respect to their different uses. The IRT analysis on the 31 tasks shows that it is possible to take a set of tasks and assemble them into a larger, multiple-item assessment. Such an assessment might be useful for evaluating curricular or instructional interventions. Using the partial credit approach, dichotomous scores on individual FKSAs can be summed to yield a total score. This total score can be analyzed using traditional psychometric approaches that are common for summative measures. Specifically, one can use IRT to model performance on the assessment as a function of student and item features. These item features (e.g., category intersections, β) can be analyzed to yield important information about the relative difficulty of items and their ability to discriminate between students at varying levels of competence. In our sample, this analysis indicated that the items discriminate best amongst higher ability students, indicating that the most precise estimates (lowest error) around students' ability occur for those students with positive theta values on the latent trait.

Although these findings about the performance of the set of items is informative, the ability to understand individual tasks' performance might be just as important for practicing teachers. In this paper we demonstrate that analyzing the frequency of credit assigned to different FKSAs can give an estimate of the relative difficulty of FKSAs as they are manifest within the item. Analyzing the correlational data between FKSAs provides further information about the observed relationships between FKSAs. The magnitude of these correlations suggests that these FKSAs are distinct, yet related, constructs. For the item we analyzed in depth, Task 26, the correlational data and FKSA frequency data together suggest that there are three potential stumbling blocks for students: (a) deciding whether the model is adequate, (b) describing features of the model that support its adequacy, and (c) deciding whether, and describing why, the model shows conservation of mass. This type of analysis shows the adequacy of the tasks for providing insights into what students know and can do.

These data also support the utility of our design approach for developing tasks and rubrics and scoring those tasks (e.g., McElhaney, 2018). The rubric components are explicitly connected to a FKSA and an associated evidence statement. This means that FKSA scores can be aggregated within an item and then used to model student performance and to calculate item and test information. This type of modeling is important for the inferential aspect of the validity argument (Pellegrino et al., 2016). Moreover, total score retains its constituent pieces and thus allows an investigation into the individual FKSAs that are demonstrated in student responses. Breaking down total scores into these component measures provides further information about the pattern of responses and insight into cognitive factors that might have produced the scores. This shows that the cognitive connections in the task design can be maintained through scoring and can then yield insights about the nature of student thought, supporting the cognitive aspect of the validity argument (Pellegrino et al., 2016). Together, these data provide further information about the tasks and their ability to provide teachers with information that might usefully inform their instruction.

Limitations

There are some limitations to the reported analysis. First, although we have analyzed data from half the assessment tasks in the physical science corpus there are still 32 tasks centered on the energy topic that we have not yet analyzed. Second, as with any study using assessment tasks, the opportunity for students to learn is an important contextualizing factor. We attempted to recruit a sample of students that would have had an opportunity to experience NGSS-aligned instruction through recruiting teachers that were using a curriculum that combines science and engineering practices with disciplinary core ideas. However, we have no record of what instruction in these classes looked like, and how much emphasis was placed on the multiple NGSS dimensions. Recall that these data were collected at the end of the 2015-2016 academic year. At that time, many of the participating districts were in their first year of implementing NGSS and the IQWST curriculum. Further, for teachers whose instruction fit the vision of the NGSS it is very likely that their classroom assessments did not. For effective multi-dimensional learning, students need instruction and assessment to be aligned. Therefore, it is unsurprising that many students only received partial credit on their responses, and that our items were difficult for students. Additionally, students in this sample had no experience with the portal before their participation in this study. That means they were not familiar with the interface nor the tools (e.g., the drawing tool, molecular workbench, etc.). If students had been more familiar with the interface we speculate that they would have completed more of their assigned assessment tasks (see Figure 1). Third, our scoring approach has many positive benefits but it is also time-intensive. We are working on a version of the rubrics that might be faster and simpler for teachers to apply, while still retaining the FKSA-design basis (Gane, McElhaney, Harris, Krajcik, & Pellegrino, 2017). Although these rubrics will likely speed teachers' interpretation of their student responses, it will still take significant time for teachers with many students. Automated analysis of constructed response data, including drawings and models, need to be explored to ease this time burden.

Conclusion

NGSS implementation requires classroom assessments that can provide information about students' multi-dimensional learning. As such, constructed response assessment tasks, which allow for partial scores on different aspects of integrated performance, are critical. These results provide evidence for inferential validity claims for our three-dimensional assessment tasks and their scoring rubrics: Task performance can be psychometrically modeled as a function of student proficiency and task features. Further, they point to the utility of using an item design and scoring model where individual features of a student response are scored separately and then summed to yield a total score. This maximizes score reliability while maintaining diagnosticity. We invite others to follow our design process (Harris et al., 2016) to create their own three-dimensional assessments and/or use our assessments in teaching and research on science learning. Widespread adoption of multi-dimensional assessments are crucial for teachers implementing NGSS-aligned instruction and for researchers measuring changes in students' proficiency resulting from educational interventions.

Acknowledgements

This material is based upon work supported by the Gordon and Betty Moore Foundation (Grant 4482) and by the National Science Foundation (NSF) (Grants 1316903, 1316908, and 1316874).

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Gordon and Betty Moore Foundation and NSF.

References

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33-63.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah: NJ. Earlbaum.

Gane, B. D., McElhaney, . W., Harris, C. J., Krajcik, J. S., & Pellegrino, J. W. (2017, March). *Classroom-based assessment tasks and rubrics: Using student responses as evidence of three-dimensional learning*. Presentation at the National Science Teachers Association (NSTA) National Conference, Los Angeles, CA.

Gorin, J. S. & Mislevy, R. J. (2013). *Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment.* Commissioned paper presented at the K-12 Center at ITS Invitational Research Symposium on Science Assessment, Washington DC.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & McElhaney, K. W. (2016). *Constructing assessment tasks that blend disciplinary core ideas, crosscutting concepts, and science practices for classroom formative applications.* Menlo Park, CA: SRI International.

Herrmann-Abell, C. F., Flanagan, J. C., & Roseman, J. E. (2013). *Developing and evaluating and 8th grade curriculum unit that links foundational chemistry to biological growth: Using student measures to evaluate the promise of the intervention*. Paper presented at the 2013 NARST Annual International Conference, Rio Grande, Puerto Rico.

McElhaney, K.W., Zaidi, S. Z., Gane, B. D., Krajcik, J. S., Alozie, N. M., & Harris, C. J. (2018). *Designing NGSS-aligned Assessment Tasks and Rubrics to Support Classroom-based Formative Assessment.* Paper presented at the 2018 NARST Annual International Conference, Atlanta, GA.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.) *Educational Measurement* (4th Ed., pp. 17-64). Westport, CT: Praeger Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

National Research Council (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council (2014). *Developing assessments for the Next Generation Science Standards.* Washington, DC: The National Academies Press.

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science, 340*, 320-323.

Pellegrino, J. W., Chudowsky, N, & Glaser, R. (Eds). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academies Press.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (Eds.) (2014). *Developing assessments for the Next Generation Science Standards.* Washington, DC: The National Academies Press.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and Item Response Theory analysis. *Journal of Statistical Software, 17*(5), 1-25.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments, Journal of Research in Science Teaching, 35, 265-296.

Tables

Table 1. District demographics for academic year 2015-2016.

District	Enrollment	Limited English Proficiency	Low Income (FRL+)	IEP	Racial/Ethnic Background
Beloit, WI	7,000	15%	78%	14%	Black 27% Hispanic 30% White 41%
Waukegan, IL	17,000	31%	73%	13%	Black 15% Hispanic 78% White 4%
Oklahoma City,OK	46,000	32%	80%	12%	Black 25% Hispanic 51% White 19%

Table 2. The seven learning performances that collectively form the “chemical reaction” topic.

ID	Learning Performance
C01	analyze and interpret data to determine whether substances are the same based upon characteristic properties
C02	construct a scientific explanation about whether a reaction has occurred using properties of substances before and after the substances interact
C03	evaluate whether a model explains that different molecular substances are made from different types and/or arrangements of atoms
C04	evaluate whether a model explains that a chemical reaction produces new substances and conserves atoms
C05	use a model to explain that in a chemical reaction atoms are regrouped and why mass is conserved
C06	develop a model of a chemical reaction that explains new substances are formed by the regrouping of atoms, and that mass is conserved
C07	evaluate whether a model explains that a chemical reaction produces new substances and conserves mass because atoms are conserved

Table 3. FKSAs for LPC04

FKSA	Statement
G	Ability to support a model evaluation using a statement that molecular substances exhibit unique atomic groupings
H	Ability to determine whether a model explains that a chemical reaction produces new substances
I	Ability to determine whether a model explains that a chemical reaction conserves atoms
J	Ability to support model use, development, or evaluation by explaining that a chemical reaction conserves atoms and/or mass.

Table 4 FKSAs for LPC06

FKSA	Statement
J	Ability to support model use, development, or evaluation by explaining that a chemical reaction conserves atoms and/or mass
L	Ability to support model use, development, or evaluation by explaining that chemical reactions regroup atoms
M	Ability to develop a model of a chemical reaction that regroups and conserves atoms

Table 5. Response count, by task ID and learning performance.

Learning Performance	Task ID	Response Count
C01	1	136
C01	3	119
C01	4	142
C01	5	128
C01	6	90
C01	33	157
C01	34	93
C02	13	161
C02	14	124
C02	15	149
C02	18	139
C02	19	58
C02	39	114
C02	80	103
C03	8	87
C03	9	139
C03	38	140
C04	26	135
C04	27	142
C04	28	157
C04	36	155
C05	23	134
C05	24	95
C05	37	136

C06	21	124
C06	22	73
C06	25	93
C06	32	73
C07	29	121
C07	30	73
C07	31	131
Total	N/A	3721

Table 6. Example full credit student response and associated coding for the Factory Reaction task (ID#: 021-03-c06).

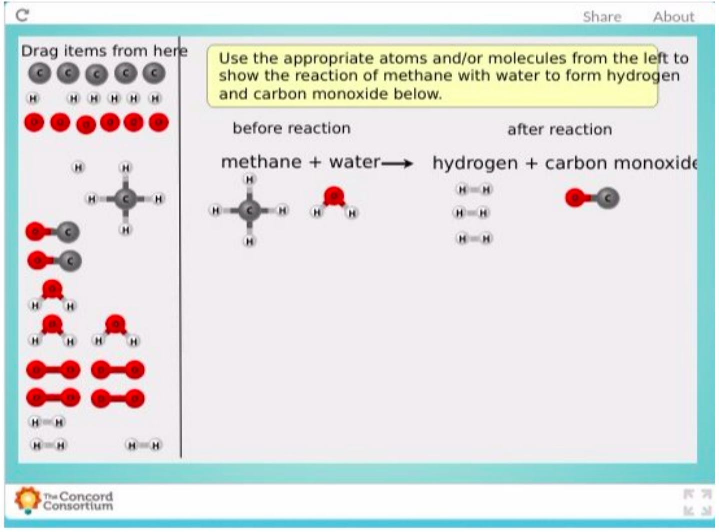
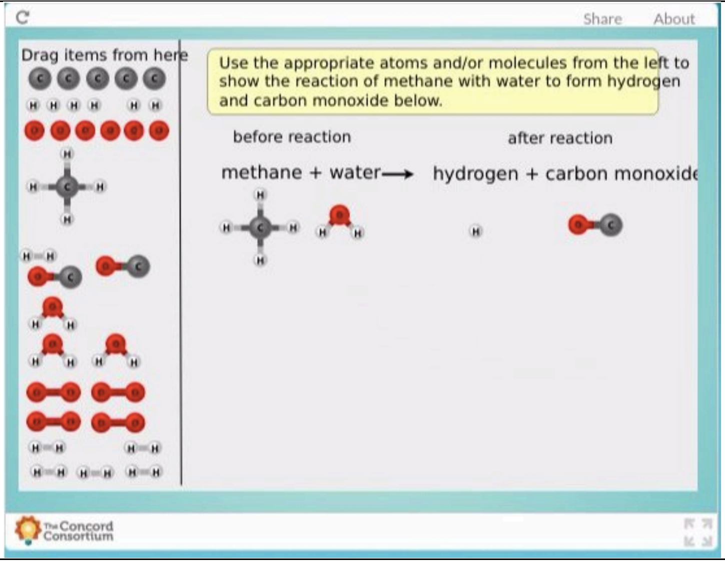
Student Response (drawing)	**Student Response (text)**
 Use the appropriate atoms and/or molecules from the left to show the reaction of methane with water to form hydrogen and carbon monoxide below. Share About Drag items from here before reaction after reaction methane + water⟶ hydrogen + carbon monoxide The Concord Consortium	during the reaction, the atoms seem to link with other atoms. as shown in the model, the oxygen and carbon atoms link together after the reaction, while before hand the were separate. as for the hydrogen atoms, they disconnected from the oxygen and the carbon atoms, taking a place for themselves away from them. as for how my model shows that the mass remains the same, the amount of atoms within the model remains the same, meaning that the mass has not changed, but the ways at which the atoms are connected have.
FKSA	**Score**
(J) Ability to support model use, development, or evaluation by explaining that a chemical reaction conserves atoms and/or mass	1
(L) Ability to support model use, development, or evaluation by explaining that chemical reactions regroup atoms	1
(M) Ability to develop a model of a chemical reaction that regroups and conserves atoms	1
Total	3

Table 7. Example partial credit student response and associated coding for the Factory Reaction task (ID#: 021-03-c06).

Student Response (drawing)	**Student Response (text)**
 Use the appropriate atoms and/or molecules from the left to show the reaction of methane with water to form hydrogen and carbon monoxide below. Share About Drag items from here before reaction after reaction methane + water ⟶ hydrogen + carbon monoxide The Concord Consortium	Five out of the six hydrogens left the methane and the water and only one was left at the end of the reaction then the oxygen and carbon atoms fused together to make carbon monoxide.

FKSA	**Score**
(J) Ability to support model use, development, or evaluation by explaining that a chemical reaction conserves atoms and/or mass	0
(L) Ability to support model use, development, or evaluation by explaining that chemical reactions regroup atoms	1
(M) Ability to develop a model of a chemical reaction that regroups and conserves atoms	0
Total	1

Table 8. Fit indices for GPCM analysis (Task 35 dropped).

Index	Rasch (slope parameter = 1)	1PL (slope parameter same for all items)	2PL* (unconstrained slope parameter)
AIC	7411.60	7413.27	N/A
BIC	7850.69	7857.09	N/A
Log.likelihood	-3612.80	-3612.64	N/A

$p = .57$

*We attempted to test the full GPCM model where each item is allowed a different slope parameter, but the model failed to converge.

Table 9. FKSA correlation matrix – Task 26

	FKSA G	FKSA H	FKSA I	FKSA J	Total Score
FKSA G	--	.11	.03	.20*	.53**
FKSA H		--	.22**	.06	.70**
FKSA I			--	.24**	.63**
FKSA J				--	.49**
Total Score					--

*p < .05; **p < .01

Figures

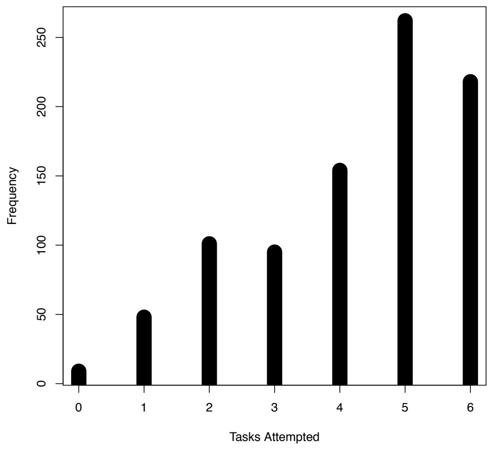


Figure 1. Number of tasks attempted. Students were assigned booklets that contained six tasks. Not all students attempted all tasks.

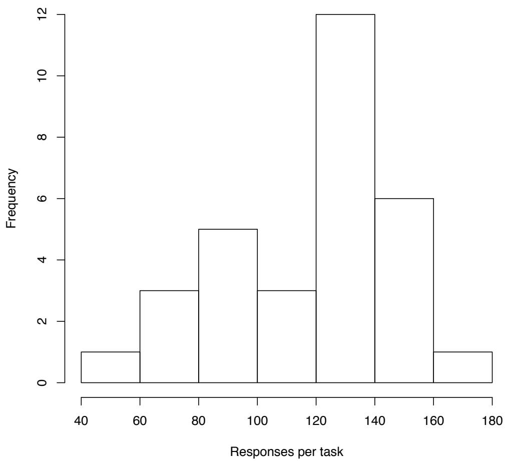


Figure 2. Number of responses per task, for the 31 tasks included in the IRT analysis.

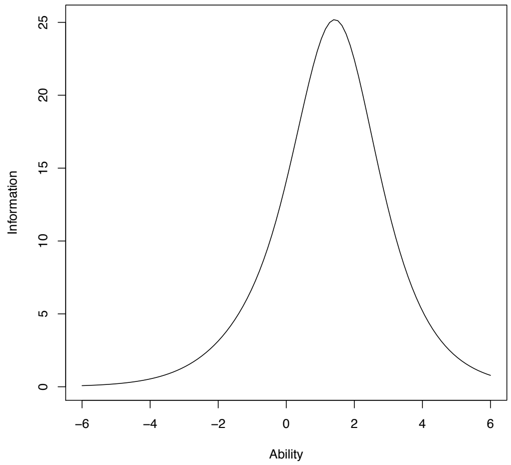


Figure 3. Test information function for 31 tasks (7 learning performances) in the chemical reaction topic. Highest discrimination occurs around θ = +1.5.

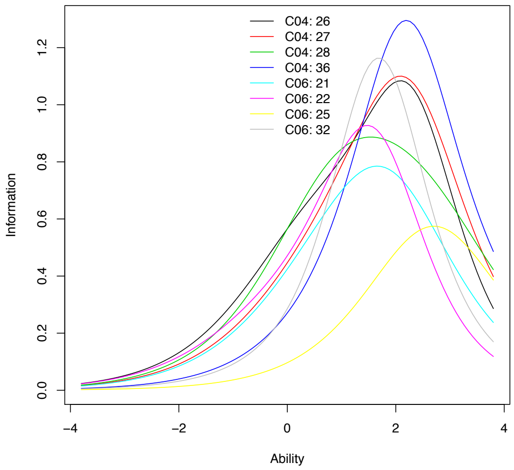


Figure 4. Item information functions for the eight items in learning performances C04 and C06. Legend indicates the learning performance and task number.

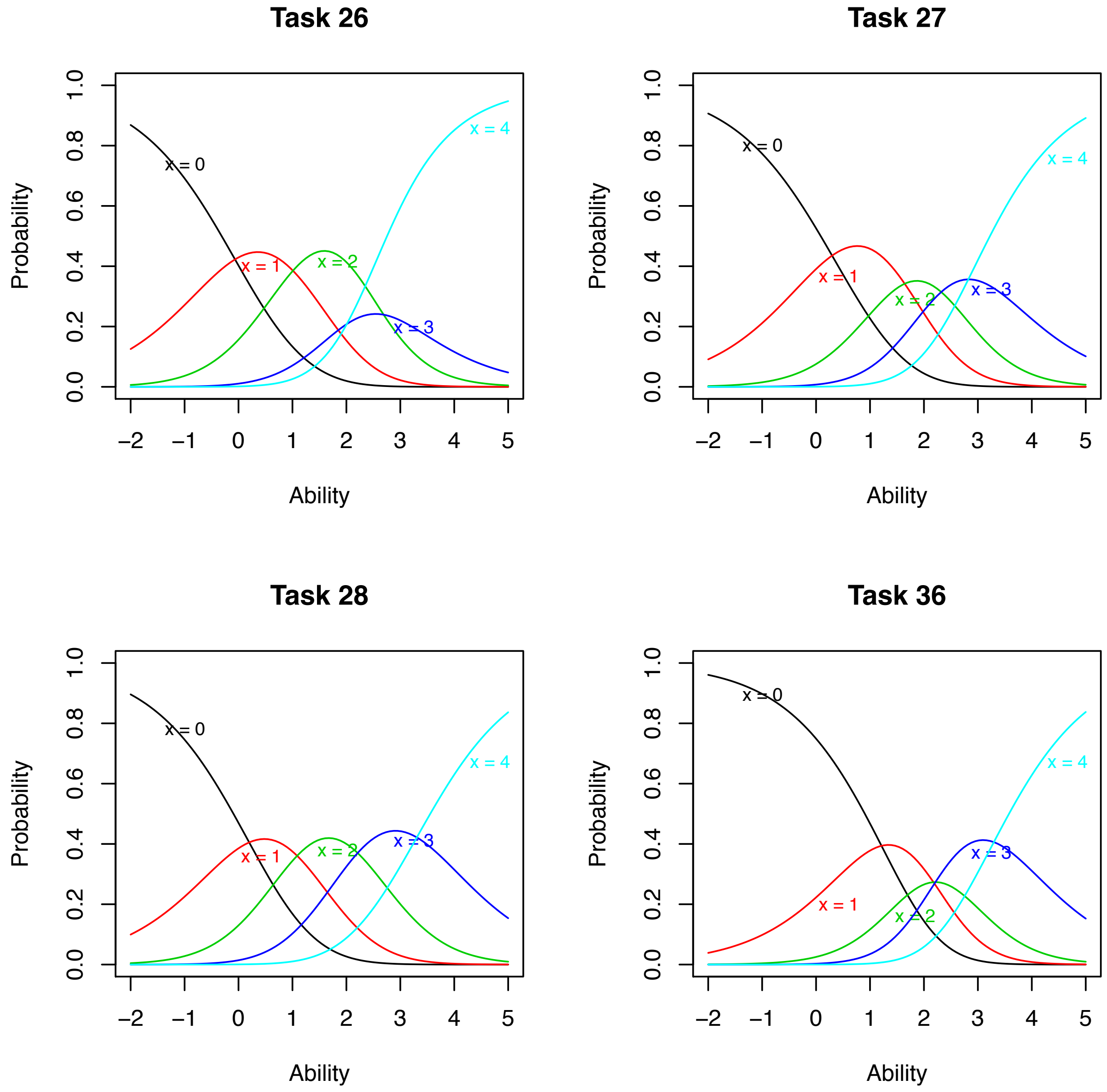


Figure 5. Category response characteristic curves (ICCs) for the four assessment tasks in learning performance C04; each function represents the probability of an examinee with a given θ earning a total score of 0–4.

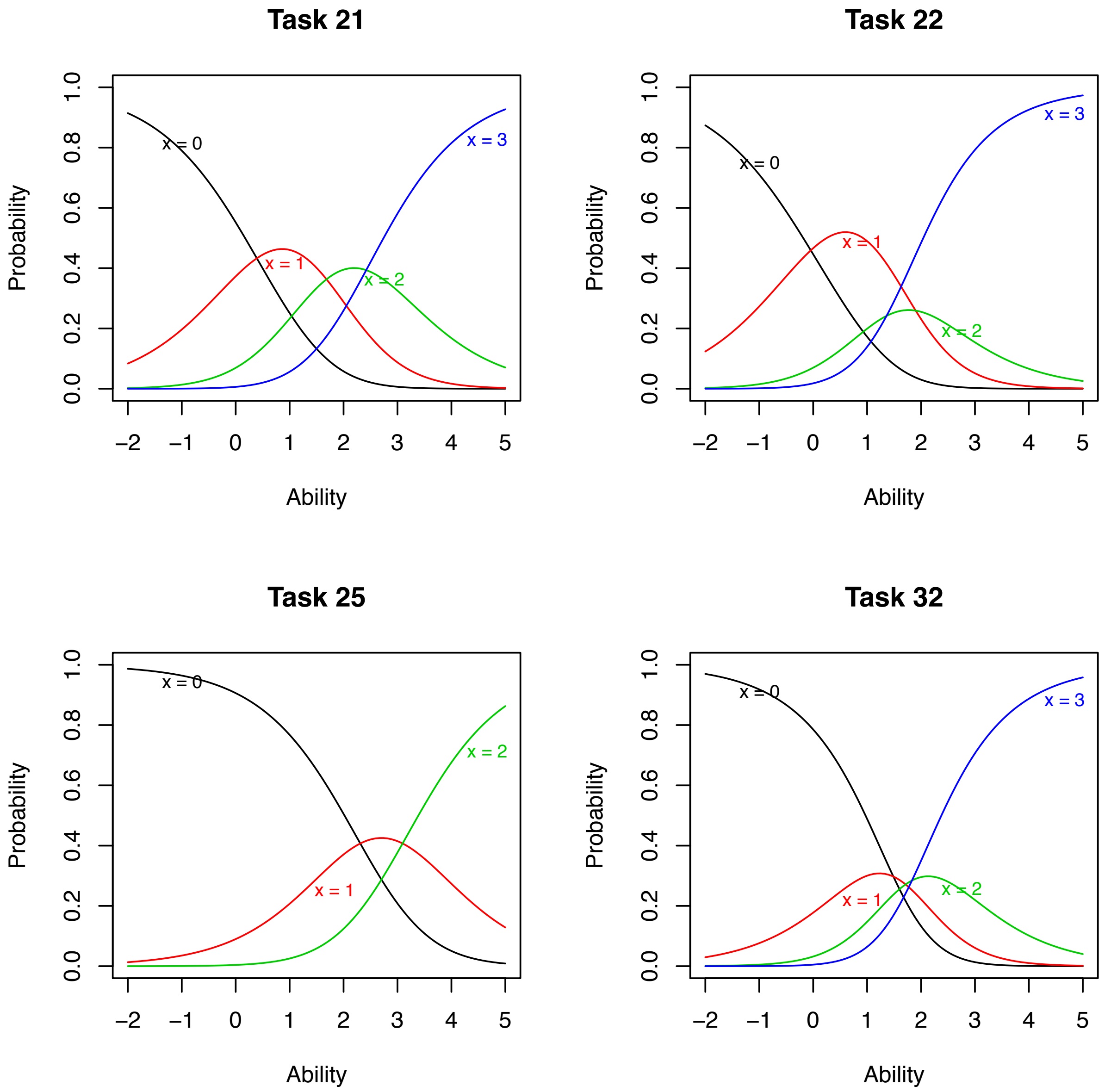
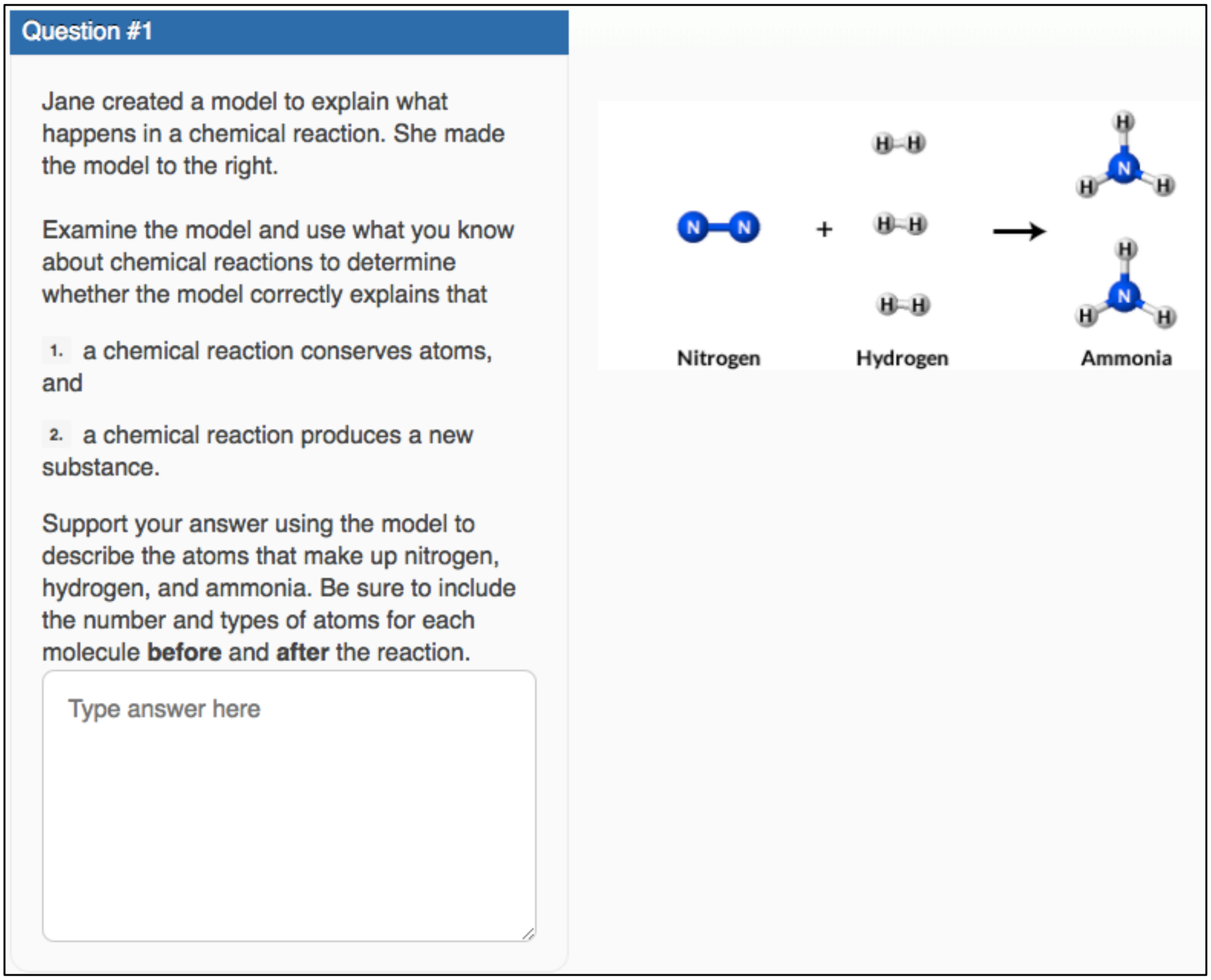


Figure 6. Category response characteristic curves (ICCs) for the four assessment tasks in learning performance C06; each function represents the probability of an examinee with a given θ earning a total score of 0–3. Task 25 has only three curves because no one in the sample received full credit (i.e., total score = 3).



Question #1

Jane created a model to explain what happens in a chemical reaction. She made the model to the right.

Examine the model and use what you know about chemical reactions to determine whether the model correctly explains that

1. a chemical reaction conserves atoms, and
2. a chemical reaction produces a new substance.

Support your answer using the model to describe the atoms that make up nitrogen, hydrogen, and ammonia. Be sure to include the number and types of atoms for each molecule **before** and **after** the reaction.

Type answer here

Figure 7. Task 26 (LP C04).

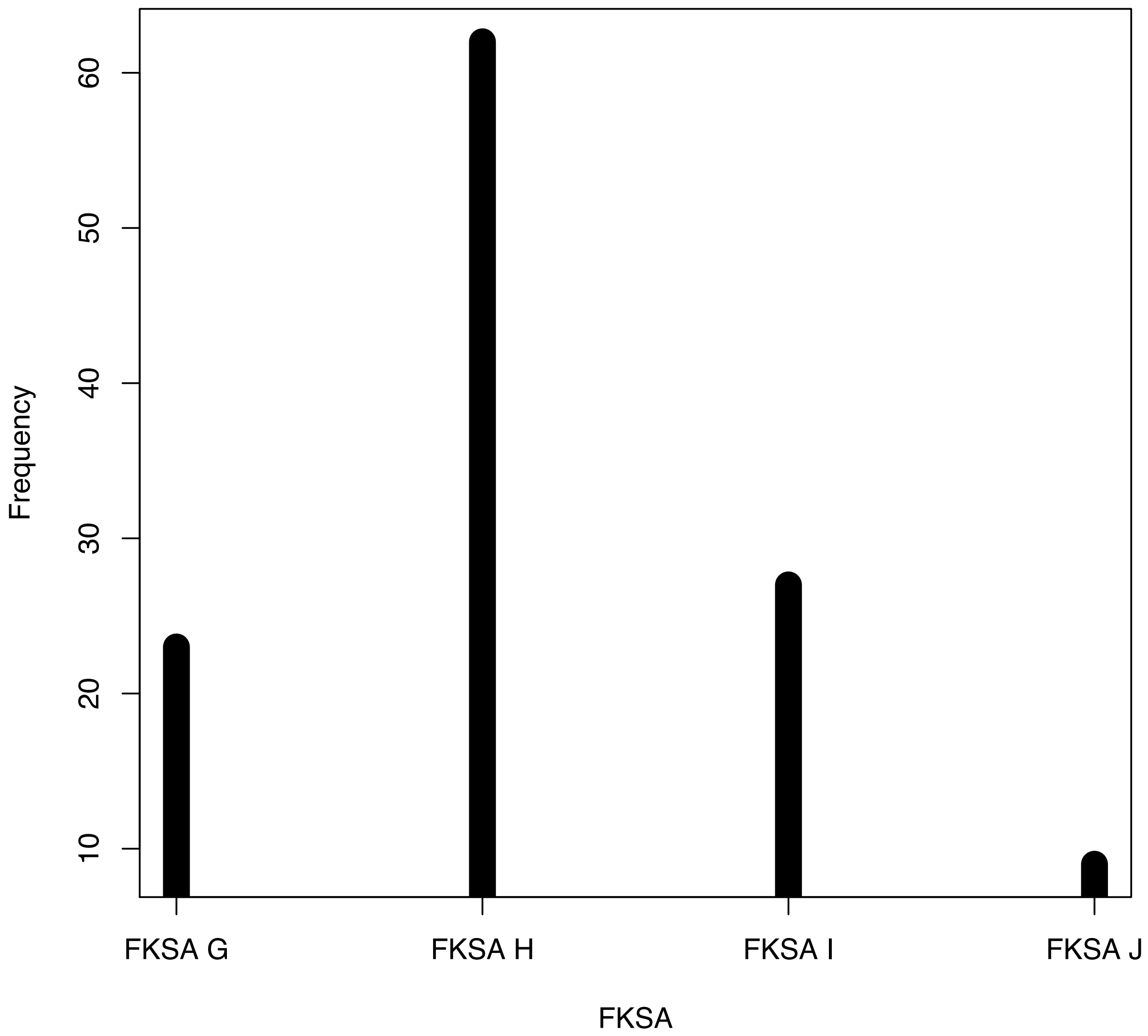


Figure 8. Observed responses by FKSA – Task 26, LPC04.