



**Research Report**  
ETS RR-13-19

# **Cognitive Interviews as a Tool for Investigating the Validity of Content Knowledge for Teaching Assessments**

---

**Heather Howell**

**Geoffrey Phelps**

**Andrew J. Croft**

**David Kirui**

**Drew Gitomer**

**October 2013**

# ETS Research Report Series

---

## EIGNOR EXECUTIVE EDITOR

James Carlson  
*Principal Psychometrician*

## ASSOCIATE EDITORS

Beata Beigman Klebanov  
*Research Scientist*

Heather Buzick  
*Research Scientist*

Brent Bridgeman  
*Distinguished Presidential Appointee*

Keelan Evanini  
*Managing Research Scientist*

Marna Golub-Smith  
*Principal Psychometrician*

Shelby Haberman  
*Distinguished Presidential Appointee*

Gary Ockey  
*Research Scientist*

Donald Powers  
*Managing Principal Research Scientist*

Gautam Puhan  
*Senior Psychometrician*

John Sabatini  
*Managing Principal Research Scientist*

Matthias von Davier  
*Director, Research*

Rebecca Zwick  
*Distinguished Presidential Appointee*

## PRODUCTION EDITORS

Kim Fryer  
*Manager, Editing Services*

Ruth Greenwood  
*Editor*

---

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

**Cognitive Interviews as a Tool for Investigating the Validity of Content Knowledge for  
Teaching Assessments**

Heather Howell, Geoffrey Phelps, Andrew J. Croft, and David Kirui

ETS, Princeton, New Jersey

Drew Gitomer

Rutgers University, New Brunswick, New Jersey

October 2013

Find other ETS-published reports by searching the ETS  
ReSEARCHER database at <http://search.ets.org/researcher/>

To obtain a copy of an ETS research report, please visit  
<http://www.ets.org/research/contact.html>

**Action Editor:** Heather Buzick

**Reviewers:** Madeleine Keehner and Gary Sykes

Copyright © 2013 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and LISTENING. LEARNING. LEADING. are  
registered trademarks of Educational Testing Service (ETS).



## Abstract

This report provides a description of a cognitive interview study investigating validity of assessments designed to measure *content knowledge for teaching* (CKT). The report is intended both to provide information on the validity of the CKT measures and to provide guidance to researchers interested in replicating the design. The study takes an argument-based approach to investigating validity by first articulating interpretive arguments that are central to the CKT measurement theory and then using the cognitive interview data to evaluate these arguments (Kane, 2006). The study is based on 30 interviews of elementary mathematics teachers and 30 interviews of elementary English language arts teachers. Teachers were selected using previous CKT assessment scores to represent high- and low-scoring groups for each subject. The cognitive interviews were conducted separately for each subject and responses were coded and then analyzed to investigate the scoring and extrapolation inferences for the validity argument. Findings strongly support the scoring inference, providing evidence that the item keying for the items is correct. Results also indicate that the participants reasoned about the item in ways that conformed with the reasoning outlined in the *task design rationales* (TDR) for each item. These TDRs represent what reasoning should look like for each of these items for a respondent drawing on the desired CKT knowledge. As such, conformity with the TDRs supports the extrapolation inference, providing evidence that the reasoning used by the participant represents the underlying knowledge and skill domain we intend to measure through CKT assessments. The study design, instruments, methods, and results are described in detail, with discussion included to support researchers interested in replicating or capitalizing on the study design.

Key words: assessments, validity, teaching, content knowledge, cognitive interview

## **Acknowledgments**

This study was conducted in the context of the *Measures of Effective Teaching Study*, supported by the Bill and Melinda Gates Foundation, and was generously supported by funding from Educational Testing Service. Many individuals contributed to the study design, data collection, data coding, analysis, and writing of this report. We would like to thank Alli Brettschneider, Claudia Guerschanik, Luis Leyva, and Barbara Weren.

## Table of Contents

	Page
Study Overview .....	2
Research Questions .....	3
Methodology .....	5
Instrument Development.....	5
Selecting Research Participants .....	8
Collecting Interview Data.....	11
Coding Research Data.....	15
Results and Discussion .....	28
Results for Research Question 1 .....	28
Results for Research Question 2.....	33
Results for Research Question 3 .....	34
Results for Research Question 4.....	36
Summary and Conclusion .....	36
References .....	39
Appendix A. Task Design Rationales for Mathematics Items.....	40
Appendix B. Task Design Rationales for ELA Items.....	69

## List of Tables

	Page
Table 1. Descriptive Statistics of MET Content Knowledge for Teaching Assessment Scale Scores .....	3
Table 2. Number of Valid Mathematics and ELA Assessment Scores by District .....	10
Table 3. District 1 Quartile Totals .....	10
Table 4. District 1 Quartiles 2 and 4 Totals (Teachers).....	11
Table 5. Conformity of Mathematics Responses to TDR by Quartile Group .....	29
Table 6. Conformity of ELA Responses to TDR by Quartile Group .....	29
Table 7. Alignment Between Correct/Incorrect Answer and Conforming Reasoning for Mathematics .....	33
Table 8. Alignment Between Correct/Incorrect Answer and Conforming Reasoning for ELA ..	34



## List of Figures

	Page
Figure 1. Mathematics task design rationale.....	22
Figure 2. Distribution of interview participants' English language arts scores.....	31
Figure 3. Distribution of interview participants' mathematics scores.....	31

An important aspect of the validity of an assessment is the quality of the connection that can be made between performance on the measure and the conclusions one would like to draw (Kane, 2006). This study utilizes Kane's argumentation approach to validity and cognitive *think-aloud* interviews to investigate the validity of two assessments designed to measure *content knowledge for teaching* (CKT). It also serves as an example of how cognitive interviews can be used in this type of validity work and draws attention to some of the key methodological choices to consider in using such a methodology.

According to Kane, establishing validity depends on specifying two arguments. "An *interpretive argument* specifies the proposed interpretations and uses of assessment results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances. The *validity argument* provides an evaluation of the interpretive argument" (Kane, 2006, p. 23). Kane further listed a number of inference types that one must be attentive to in specifying the interpretive argument; these include scoring, generalization, extrapolation, and decision making/implication.

This study focuses on validity evidence for two of these inferences. The first is the scoring inference, which specifies that items are keyed accurately and in ways that capture only construct-relevant variance. The second is Kane's (2006) extrapolation inference, or the degree to which the reasoning demanded by the assessment items approximates the reasoning that would be demanded in the situations about which one would like to draw conclusions. In this study, we were interested in how teachers reason about the recurrent content problems that occur in teaching practice. One piece of evidence for the validity of CKT measures is the extent to which teachers' reasoning about these items, situated in tasks of teaching, approximates the reasoning they would use in actual teaching. This is the type of reasoning that is intended by the item design and is embedded in the design theory for each item.

We examined performance on the Measures of Effective Teaching (MET) elementary CKT assessments by conducting cognitive interviews with a set of teachers who had previously taken the assessments. The interview data support two hypotheses. The first is that the item keying for the 20 items selected is correct, lending support to the scoring inference. There is little construct-irrelevant variance and few instances of defensible reasoning that support an answer other than the intended key. The second hypothesis is that participants reason about the items as intended, where *as intended* refers to the conformity of demonstrated reasoning to the reasoning

that is specified in the task design rationales (TDR) for each item. These TDRs, built on careful expert review, represent our best image of what the use of such reasoning should look like if the construct successfully represents the knowledge demands of teaching. As such, conformity with them can be taken to support the extrapolation inference that reasoning used by the participant represents a use of the underlying knowledge and skill domain that we intend to measure through CKT assessments.

This report provides a description of the research questions, study design, and research methods used to conduct the validity study. We have included in the report a careful description of the design theory for the CKT assessments that we set out to validate. We think that the design of this study has applicability beyond just CKT and can inform other groups interested in validating item design theory.

### **Study Overview**

This study was conducted in the context of the MET project. MET was a large-scale study designed to investigate the relationships among four different types of measures associated with teaching and teacher quality: student achievement or value-added measures for classrooms, observation measures of classroom instruction, student evaluations of their classroom environment, and teacher knowledge assessments. Five assessments of CKT were developed for MET, including Grades 4-5 mathematics, Grades 6-8 mathematics, Algebra 1, Grades 4-6 English language arts (ELA), and Grades 7-9 ELA. The MET study was conducted in six school districts. Data were collected on each participating teacher for each of the major measures and school environment.

The cognitive interview study is nested within the larger MET study and leverages the full sample for the elementary mathematics and ELA assessments. The scored sample consisted of 681 individual teachers who completed 952 assessments, 271 of whom completed both a mathematics and ELA assessment. These assessments included mostly selected response items and a small number of constructed response (CR) items. The selected response items included two types: multiple choice (MC) items that required participants to select a correct choice from four options and table items that included a series of yes or no questions. The Grades 4-5 mathematics form contained a total of 40 items, and the form reliability was 0.76. The Grades 4-6 ELA form contained a total of 53 items and the form reliability was 0.74. Both

reliabilities are reasonable for assessments of this length. Performance statistics are shown in Table 1.

**Table 1**

***Descriptive Statistics of MET Content Knowledge for Teaching Assessment Scale Scores***

Assessment	n	Min%	Max%	Mean%	SD
Grades 4-5 mathematics	397	20.5	93.2	52.2	14.4
Grades 4-6 ELA	555	30.8	89.4	66.4	11.7

Note. MET = Measures of Effective Teaching; ELA = English language arts.

Nesting the study within a pre-existing sample that included existing assessment scores provided a useful basis for selecting items to include in the cognitive interview study and for selecting participants who differed in overall score performance. This is a study design that is available in many test validation contexts, where extant data, in either pilot or operational form, can be used as a basis for identifying target groups of participants based on their score range and for purposely selecting participants for the cognitive interview validation study.

### **Research Questions**

The study was designed to examine four primary research questions focused on evaluating the scoring validity argument and the extrapolation validity argument for the CKT assessments.

1. To what extent do teachers' classifications on assessment score (MET) correspond to their classifications on conformity with the intended reasoning?
2. For each item and for all items, to what extent is correct and incorrect reasoning associated with correct and incorrect answers, respectively?
3. For responses for which correct/incorrect reasoning does not associate with correct/incorrect answers (at both the item level and for all items), to what extent and in what ways are these responses due to:
  - a. defensible reasoning that supports a different item key?
  - b. nondefensible reasoning that is associated with the correct key?

4. To what extent do the items remind teachers of something they have experienced in their teaching? To what extent do teachers perceive the items to be authentic problems that would be encountered in teaching?

Research Questions 2 and 3 address Kane's (2006) *scoring* inference. "The scoring inference employs a scoring rule to assign a score to each student's performance on the test tasks. For MC tests, the scoring rule consists of an answer key for the test.... The scoring inference relies on two basic assumptions: that the scoring criteria are reasonable and that they are applied appropriately" (Kane, 2006, p. 24). The second research question evaluates whether the participants reasoning and knowledge is associated with correct and incorrect answers, thus providing support for the claim that the scoring rules and overall score scales are defensible. The third research question examines the nature of the responses that do not fit the scoring rule for evidence of incorrect keying and/or unanticipated alternate defensible solutions.

Research Questions 1, 2, and 3 address Kane's (2006) *extrapolation* inference: "The extrapolation inference assumes that the test tasks provide adequate measures of the competencies of interest (those developed in the courses) and are not overly influenced by extraneous factors (e.g., test format)" (Kane, 2006, p. 24). Extrapolation includes the inference that the portion of the target domain that can be measured extrapolates to the larger domain of interest, in this case, that the CKT that is measurable on this type of test extrapolates to the CKT construct more broadly. It also includes the inference that what is measured is related to the knowledge and skills of the construct and is not overly subject to construct-irrelevant variance. Research Question 1 focuses on the extrapolation of scores on the MET assessments to the target score (the true score that would represent an individual's CKT) by looking at the relationship between MET assessment scores and reasoning coded as representing CKT. Reasoning coded as representing CKT is used here as a proxy for the target score, although often a criterion measure serves this function if such a measure is available. Research Question 2 focuses similarly on extrapolation but at the item level. Both of these research questions provide analytic evidence for the extrapolation inference, focusing on, as Kane suggested, "general notions about overlap in the processes employed in responding to the test tasks and other tasks in the target domain" (Kane, 2006, p. 35). Question 4 provides basic confirmation that the test design has succeeded in presenting testing tasks that teachers perceive as directly related to the types of content problems teachers encounter in their work. Because part of the theory of CKT is that the measured

knowledge is the content knowledge that would be used in teaching, these reports from participants provide evidence supporting an important feature of the CKT assessment design theory.

## **Methodology**

Below we describe the methods used to design the study instruments, select research participants, collect interview data, and code interview data.

### **Instrument Development**

**Selecting items for interview test forms.** Because cognitive interviews are time intensive, they often require selecting a subset of assessment items from a test or, even more broadly, as a sample representing an assessment type. Typically, the goal is to select items that are in some sense representative of the larger domain and thus provide general insight into the performance or characteristics of this larger set of items. While this is rarely done through systematic sampling, the design logic almost always depends on drawing inferences from the actual items included in the cognitive interviews to the more general class of items. For this cognitive interview study, we were interested in selecting items that we felt represented the larger CKT domain. Selection criteria included the following:

- Select better performing items if possible (as measured by higher biserial correlations). Our goal at this stage of the project was to understand our *best* efforts in designing CKT items—we wanted to understand what works so that we can use this as a model for future development.
- Exclude items with known design issues unless there is a strong warrant for inclusion. Inclusion of items with known design issues was likely only to confirm what we already knew about these items and was not a priority at this stage in the project.
- Include a range of difficulty level among items (as measured by the percentage of assessment-takers who answered the item correctly during the MET administration).
- Include items representing a diversity of content topics.
- Include items representing a diversity of tasks of teaching.
- Include items representing a diversity of knowledge domains.

- Include a variety of item types (CR, table items, and MC).

To provide a basis for probing on item specific features, an initial sort was completed, taking into account the above criteria, and grouping items as *best/good/OK/fair/poor*. To the extent possible, two project staff members for each of the subject areas reviewed available items independently and then worked together to make decisions to help ensure that items were appropriately sorted. Given the number of criteria used to select items, this process needed to be somewhat flexible and iterative. This initial sort was brought to the entire project team (for each subject area) for review and revision, and initial items were selected for inclusion in pilot interviews. Thirteen ELA items were selected and 14 mathematics items were selected, with the anticipation that, based on the results of pilot interviews, we would reduce the number of items to be included in the final interview forms.

During pilot interviews, detailed notes were taken and items were timed. Based on timing, it was determined that three items in ELA and four items in mathematics would need to be cut from the list in order to achieve our goal of a cognitive interview lasting no more than 90 minutes. Each subject-area team worked collaboratively to determine which items produced the least useful data during the pilots, and the remaining items were assembled into the final interview test forms. Final interview test forms for each subject area are included in a set of supplementary materials available on request from the authors.

**Task design rationales.** A crucial tool for this study was the set of documents we call TDRs. For each item, a TDR provides an explicit rationale for the task, the hypothesized logic for the selection or construction of a correct response, and a set of hypotheses for why incorrect responses should not be selected. Each TDR acts as a design map for the individual item, filing in the item-specific component of the interpretive argument by articulating the anticipated reasoning that would lead to a correct answer. A complete set of TDRs had previously been drafted and those selected for use in the cognitive interviews received two additional cycles of revision and review. (For an example of a TDR, see Figure 1.)

**Interview protocols.** Design of the interview protocols included the design of an overall protocol for all interviews and the development of item-specific interview prompts for each item. The overall protocol was developed by the project team over the course of 2 full-day sessions. Interview questions were selected to support the research questions we hoped to answer, also taking into account the time constraint, the need to minimize learning effects during the course of

the interview, and the need to carefully craft follow-up prompts that would maximize the usefulness of responses while minimizing measurement noise to the greatest extent possible. A review of audio files from previous cognitive interviews was used to evaluate the types of responses these questions had elicited. The protocol was adjusted to avoid questions similar to those that had elicited poor responses on past interviews and to incorporate more follow-up questioning in places where pilot interviews elicited incomplete responses.

The general protocol was then customized to include item-specific prompts and responses reflecting the anticipated patterns of reasoning represented in the TDRs and targeting areas for the interviewer to follow up on depending on the responses received during the interview. The cognitive interviews represented our first opportunity to put these explicit claims to the test and to investigate whether teachers do, in fact, reason about the items in the ways that we anticipate. The interview final-interview protocol shell (without item-specific prompts) is shown here:

1. What was your answer on this item?
2. Did the scenario presented remind you of something you've experienced in your teaching? Can you help me understand what it reminds you of? [if not] Do you think this is a scenario that other teachers might encounter?
3. You selected \_\_ as your answer. Can you say why you decided this was the best answer? [Item-specific follow-ups included here.]
4. Let's go through the other options. Why did you not select \_\_ as your answer? [Item-specific follow-ups included here.]
5. Was there anything in the question that you felt was unclear? -or- You mentioned before that you felt \_\_\_ was unclear. Can you say more about that? How was it unclear? What assumptions did you end up having to make to answer the question?
6. Looking back at the context (and by that I just mean everything prior to the answer options that I was just asking you about), can you think back to when you first read this and tell me anything that you noticed, anything that jumped out at you?

**Pilot interviews.** Four pilot interviews were conducted in each subject area. These small-scale pilots were used to inform instrument revision and were also used as an opportunity to train project staff in the techniques of semistructured interviewing as discussed below. A convenience



sample of eight elementary-level teachers was recruited for the pilot. No specific selection criteria beyond teaching experience in the desired subject area at the desired level were applied.

Each pilot interview was followed by a debriefing session of the subject-specific team, and, after each interview, the general and specific protocols were revised based on the results of that discussion. (Sample protocols and other study instruments are available on request from the authors.) Sample protocols for mathematics and ELA is included in a set of supplementary materials available on request from the authors.

### **Selecting Research Participants**

**Design logic for selection.** Cognitive interview studies are often designed using participants who form a subset of a larger test taker population. When selecting the interview study participants, it is important to consider the purpose of the study and then to identify any characteristics of the larger population that need to be represented by the study participants. One way of selecting participants for inclusion in cognitive interviews is to attempt to select a group that similarly represents this population. However, participants might also be selected purposefully to focus on just one aspect of the tested population. For example, it might be important to include relatively rare characteristics of the participant population and, in this case, participants would need to be oversampled for this characteristic to ensure that it is present to be observed and studied in the cognitive interview sample. It might be necessary or desirable to exclude certain characteristics to reduce or control for variation in order to allow study of other characteristics of interest.

For this study, we had the potential to select participants from the full pool of MET teachers who had been scored on the elementary mathematics and ELA assessments. Our first decision was to select participants from a single district. In part this was simply a practical decision given the complications of securing human subjects permission from all districts on a relatively short timeline. By choosing a single district, we were also able to isolate district-level contextual factors, such as district standards and mandated curriculum materials. Because our items are, by their nature, susceptible to contextual factors such as curriculum, we judged that it was more important for our purposes to control this source of variance in ways that would allow us to observe unique contextual influences and less important to obtain a representative sample across multiple school districts. Having made this decision, we selected a district from among

those available, largely based on sample size available, in order to maximize our chances of full recruitment. This process is described in more detail in this report.

Our second decision was to select on the basis of performance on MET CKT assessments. We still had a number of choices to make—whether, for example, to sample from the extremes, to choose participants deliberately distributed by performance, and whether to use any kind of grouping in our selection process. We decided to use quartile grouping, using the quartiles from the overall MET score distribution, and to select such that half of our interviewees were high performers (Quartile 4) or lower performers (Quartile 2). This sampling method supported several goals. One was to raise the likelihood that we would have sufficient variation in our sample to ensure that we would have responses to analyze for high and low performers. A second was to allow us to compare the cognitive interview responses for distinct groups based on their overall assessment score to confirm our hypotheses that the overall assessment represents identifiable differences in actual CKT as demonstrated in the interviews.

The choice to focus on Quartile 2 rather than Quartile 1 as the low-performing group was informed by a number of considerations. Within the selected district, there were relatively few teachers in Quartile 1. Given that very low performers are less likely to participate in such studies, we anticipated that selecting Quartile 1 would have made recruitment difficult and would have reduced our sample size. Another consideration was the relationship of Quartile 1 scores to chance. Scores in the Quartile 1 range were approximately the same as scores one would expect to achieve by chance alone, and we suspected that this group might include a significant number of participants who had not engaged fully with all the items or who might have guessed frequently or rushed through the assessment. We wanted to interview participants who had made a good faith effort completing the assessment. Because our study focuses on teacher reasoning, we wished to look at low-performing cases where there was a good chance that teachers giving an incorrect answer had thought about their response rather than simply guessing.

**Narrowing to a single district.** Once we had decided to narrow our recruitment pool to a single district, we needed to select from among the six districts that had participated in the MET study. Teachers whose scores had been judged to be invalid and who were excluded from the final MET assessment analyses were also excluded from the participant pool for this study. Because we were conducting separate interviews for each subject, we examined the data separately for teachers who had tested in both subjects. Table 2 presents the number of valid

scores in each subject for each district in the study. We choose District 1 as providing the largest sample from which to select participants for both subject areas.

**Table 2**

*Number of Valid Mathematics and ELA Assessment Scores by District*

District	Grades 4-5 mathematics	Grades 4-6 ELA
1	104	131
2	95	125
3	85	141
4	19	13
5	94	116
6	n/a	29

Note. ELA = English language arts.

Next we calculated score quartiles relative to the entire MET sample for each assessment. This means that District 1’s teachers’ quartile placement was relative to how teachers performed in the entire scored sample, including the other five districts. Table 3 shows the number of District 1 teacher scores that fall into each quartile for the Grades 4-5 mathematics and Grades 4-6 ELA assessments. The District 1 population performed slightly better than the MET population as a whole, as can be seen in Table 3. However, the distribution allowed for the potential recruitment of 15 teachers from each of Quartiles 2 and 4 in each subject area, which was not the case in other districts, making District 1 the best choice for meeting recruitment targets.

**Table 3**

*District 1 Quartile Totals*

Quartile	Mathematics	ELA
1	15	24
2	21	34
3	35	34
4	33	39
Total	104	131

Note. ELA = English language arts.

About midway through recruiting, it became clear that we would not be successful in recruiting 15 teachers from the smallest quartile group pool (Mathematics Quartile 2). At that time, we examined the list of all math-eligible participants with scores in Quartiles 1 and 3 but whose scores were numerically close to Quartile 2 scores. A total of four teachers from Mathematics Quartile 1 or 3 but whose scores were numerically close to Quartile 2 scores consented to participate in the study. One additional teacher was recruited during data collection to replace a mathematics interview for which there were missing data due to interviewer error; in all, 31 mathematics interviews were conducted.

Because we were interested in looking at any evidence that would help us understand reasoning across subjects, we sought to include all participants meeting our search criteria in both subjects. A total of 18 teachers who returned consent forms qualified for and completed cognitive interviews in both subject areas. Table 4 summarizes the cross-tabulation of Quartiles 2 and 4 across subject areas.

**Table 4**

*District 1 Quartiles 2 and 4 Totals (Teachers)*

Quartile	Mathematics Q2	Mathematics Q4
ELA Q2	4	3
ELA Q4	3	8

Note. ELA = English language arts.

### Collecting Interview Data

**Interview schedule.** The cycle for each interview was as follows:

- Assigned interviewer contacted the interviewee to arrange mutually acceptable time.
- Interviewer sent the instrument to the interviewee 3 days in advance.
- Interviewer confirmed interview date and time and receipt of instrument 1 day in advance.
- Interviewer conducted interview by telephone at the arranged date and time.

- Immediately after concluding the interview, interviewer uploaded an audio file and completed a field-note file for the interview.
- Project coordinator verified the audio within 24 hours and backed up and archived all data.

Each interview conducted required approximately 2.25 hours of the interviewer's time, including time for communication, follow-up, recording field notes, and uploading all data. Each interview required 15 minutes of the project coordinator's time for the purpose of verifying and backing up files. Data collection accounted for approximately 150 hours of project staffing time total and resulted in approximately 90 hours of recorded audio data.

It is important to note that these interviews were conducted by phone, rather than face to face. There were some advantages to conducting the interviews by phone, including the ease with which other team members could sit in during training without creating an intimidating environment for the participants. There were also disadvantages. For mathematics in particular, not being able to see what the participant was writing during the interview made following the conversation a more difficult task for the interviewer. However, the main reason for conducting the interviews by phone rather than in person was simply to maximize the number of interviews we could conduct within the project's budget. Because our participant pool included teachers in six possible districts, none of which were local to ETS, the travel costs associated with in-person interviews would have been significant, and it would have been necessary to reduce the number of interviews conducted proportionally to the time constraints this would have created.

A key decision made in defining the interview process was how far in advance to send the interview test forms to participants. The decision to send the interview test forms in advance was made based on prior experience in the pilot interviews and in other similar types of interviews. Because the scheduling of pilot interviews was difficult and opportunistic, participants did not always complete the items in advance, even when the items were provided to them in advance. Our experience was that our data were of lower quality for the items the participant had not worked through in advance and, in particular, that the burden of reading and responding to the item for the first time, coupled with anxiety about doing this in front of the interviewer, tended to distract the participant and make it more difficult for the participant to articulate his or her thoughts. Because the interviews were conducted by phone, prompting the

sharing of thoughts was more difficult in these cases, and we had no opportunity to engage in preactivities that would have accustomed the participants to the experience of thinking aloud in the ways that would make up optimal study design (Clement, 2000). It was also our experience that we did not observe widespread evidence of participants manufacturing explanations for their work, which can be a concern in conducting retrospective think-alouds (Ericsson & Simon, 1985). Generally, participants' notes from their work were quite complete and they were able to report without hesitation what their thinking had been in arriving at an answer. In cases where participants encountered a portion of the item that they had not fully examined in advance, the tendency was to stop and think through an answer, rather than to supply a quick answer. There were exceptions to this trend, but we judged that these exceptions would be less damaging to the data than the types of problems that arose when participants worked through items for the first time at the time of the interview. The decision to allow 3 days was made simply in deference to the busy schedules that most teachers have. In many cases, the interviewer was able to send the instrument only 1 or 2 days in advance after speaking to the participant to find out what time he or she would be setting aside to look over it, but this was not something we could guarantee. While, in general, concurrent think-alouds are considered more useful than retrospective ones for revealing decision processes, the summary statements provided were more than adequate to our analytic needs. In fact, a strength of retrospective protocols is their tendency to bring out more statements about the final choices made; this was of key interest in this study (Kuusela & Paul, 2000).

**Training research staff.** The project teams in both mathematics and ELA were made up of ETS staff, external consultants, and project collaborators from Rutgers University. A rigorous training procedure was put in place. The purpose of such training, along with the item-specific interview protocols, was to improve reliability of interview data by ensuring that the responses and probes used would be as uniform as possible across subjects and interviewers (Clement, 2000).

Pilot interviews as noted above were conducted in part to facilitate training of less experienced staff by providing both examples of interview technique and opportunities for practice and feedback. The first pilot interview in each subject area was conducted by more experienced researchers and the debrief session focused on the choices that were made to follow up or not follow up at specific moments, the types of information that were elicited, and how this

information gathering was accomplished. Because the interviews were conducted by telephone, training also focused on how to create a welcoming environment, minimize participant discomfort, and keep the conversation going without the benefits of eye contact. Starting with the second pilot interview in each subject area, less experienced researchers were scheduled to lead portions of each review. Feedback was provided during debriefing sessions and also provided one-on-one after each session.

Each subject-matter team also met for additional training in which the group reviewed selections from the pilot interviews and discussed how the questions posed by the interviewer did or did not support access to the information needed to address the research questions. This training also included detailed instruction in the procedures for contacting interviewees, scheduling interviews, using recording equipment, and handling various potential difficulties. Additionally, the project coordinator met individually with each member of the team to perform a dry run in which 5 minutes of audio were recorded, uploaded, and checked for audio quality, and team members were not cleared to begin interviews until this dry run was complete. In some cases, this process was iterated in order to achieve optimal audio recording quality if the interviewer was using a cell phone or if the recording device required different settings. Team members with limited interview research experience continued to observe more experienced researchers as interviews commenced and were shadowed by more experienced researchers for the first two to three interviews that they conducted.

**Data management.** As described above, interviewers were required to upload data immediately after the conclusion of the interview and these files were verified and backed up to multiple locations within 24 hours. On two occasions, there were data collection issues associated with the interviewer failing to activate the recording device at the appropriate time in the interview. In the first case, the interviewer noted the error within a minute of the interview's start and asked the participant to repeat the information. In the second, the interviewer noted the error about 5 minutes later. Immediately after the interview, the interviewer recorded detailed field notes accounting for the missing 5 minutes.

All audio files were within acceptable parameters for audio quality. Audio files were professionally transcribed and returned to ETS in batches. Project personnel then cleaned the transcribed files. Cleaning of files took (on average) 3.5 hours per file including a final review, accounting for approximately 215 hours of project personnel time. Transcripts were uploaded

into the Dedoose qualitative analysis software (SocioCultural Research Consultants, 2012) and prepared for analysis by coding descriptors (interview ID code, associated scores, teacher background information, etc.) and marking sections corresponding to associated item responses.

### **Coding Research Data**

**Coding method.** The unit of analysis for questions involving the answer given or the justification given (Protocol Questions 1 and 3) was one person's response to one item, where each MC item, each CR item, and each row of a table item was considered a single item. This yielded a total of 21 items for mathematics (one CR, six MC, and three table items with multiple rows) and 18 for ELA (eight MC and two table items with multiple rows) for a total of 640 data points in mathematics and 540 in ELA. For analyses involving Protocol Questions 2, 5, and 6, there were fewer data points because table items were treated as single items rather than rows. For these analyses there were a total of 303 data points in mathematics and 300 in ELA.

Coding teams were divided by content area, with members of each content team bringing prior expertise in teaching or in the content area or in both. Two project directors coded across content areas to monitor consistency in code application, and approximately 33% of all responses were double- or group-coded and reconciled to maintain consistency of coding over time. Once fully coded, the data were exported and brought into SPSS for quantitative analysis.

Responses to each interview question were coded on a number of characteristics. Responses were coded as correct or incorrect and also flagged for explicit uncertainty in answering. They were also coded with respect to whether any part of the item caught their attention or was confusing and whether the item reminded the interviewee of teaching. The coding for responses that reminded the interviewee of teaching also differentiated between reminders that were (a) the work of teaching (as referenced in the item) and (b) content/curricular areas.

**The coding schema.** The codebook, with detailed definitions and criteria for inclusion, is included in a set of supplementary materials available on request from the authors. Coding and reconciliation for the total set of data accounted for approximately 600 hours of project personnel time. The code tree we used is as follows:

1. What was your answer?
  - Correct answer



- Incorrect answer
  - Expressed uncertainty about answer
  - Other answer (not correct/incorrect)
2. Did this remind you of your teaching?
- Answer to question
    - Yes, reminded of teaching
    - No, did not remind, but yes for other teachers
    - No, did not remind of teaching
  - Nature of reminder
    - Task of teaching
    - Not task of teaching
      - Reference to things students struggle with
      - Content/curriculum/context
    - Other
3. Justification for selection
- Justification conforms to TDR
    - Changes earlier answer to conform to TDR
    - Response to distracters contradicts conformity
    - How it conforms (criteria created for each question)
  - Justification diverges from TDR
    - Incorrect content
    - Explicit guessing/confusion
    - Does not attend to a critical aspect of the item
    - Justification is not a justification

- Reasons through contrast with nonselected
  - Misreading the item
  - Works on different task of teaching/answers different question
    - Ignores task of teaching
    - Redefines task of teaching
    - Distracted by content from task of teaching
    - Other
    - Other way of diverging
4. Construct-irrelevant confusion
- No answer to question
  - Confusion
  - No confusion
5. What jumped out [caught the participant's attention]
- Nothing jumped out
  - Something jumped out
    - Realistic for students
    - Details related to task of teaching
    - Implications for teaching practice
    - Content
    - Initial confusion
    - Realistic/authentic to the test taker
    - Way to improve the item/take it further
    - Something else
6. Other/global codes

- Egregious leading question
- Defensible non-TDR reasoning
- Noteworthy

The most important codes were those characterizing the nature of the justification for the selected option, which was coded as conforming or not conforming to the item rationale/TDR. We see this as one of the most important features of this study and also an important contribution to other studies seeking to investigate claims about a test-design theory. Detailed specification of the intended function of items provides a way to evaluate participant responses and, by extension, to build a validity argument around what the items are measuring. In Figure 1, we provide an example of a TDR for a mathematics item.

### Assessment Question

To assess her students' prior knowledge about evaluating arithmetic expressions, Ms. Santiago assigned a worksheet of problems. She noticed that Alexis answered the first two incorrectly and the next two correctly.

$$1) 7 \times 2 - 6 + 3 = 5$$

$$2) 9 - 5 + (16 \div 8) = 2$$

$$3) 9 + 24 \div 3 - 1 = 16$$

$$4) 17 - (3 + 7 \times 2) = 0$$

Which of the remaining problems is Alexis likely to answer incorrectly?

Option 1:  $8 + 7 - 12 \div 3$

Option 2:  $13 - 3 \times 2 + 5$

Option 3:  $(27 \div 3 - 4) + 8$

Option 4:  $(16 - 12) \times 5 + 10$

### What is this assessment task asking?

Although this assessment task asks you to identify which problems the student is likely to answer incorrectly, the primary challenge is figuring out what Alexis is doing based on the work samples that are given. This means first figuring Alexis's source of confusion as demonstrated by the combination of correct and incorrect work. The next step is determining how her confusion might lead to answering incorrectly for each of the four problem choices. Answering this assessment task is aided by the knowledge that students

who have learned to solve problems drawing on mnemonics used to help them remember order of operations are likely to make the same errors that Alexis demonstrates in her work.

### What information is important?

To answer this assessment task, you first need to analyze the four examples of Alexis's work. You need to understand what she did to get the first two problems wrong and then test your hunch about her confusion to see if it is consistent with answering the other problems correctly.

In the first example, Alexis should have done the following:

$$\begin{array}{ll} 7 \times 2 - 6 + 3 & \text{multiply 7 by 2} \\ = 14 - 6 + 3 & \text{subtract 6 from 14} \\ = 8 + 3 & \text{add 8 and 3} \\ = 11 & \end{array}$$

However, Alexis gave 5 as an answer instead of 11. What might she have done that would explain this?

One way of combining the numbers incorrectly that leads to a result of 5 is shown below:

$$\begin{array}{ll} 7 \times 2 - 6 + 3 & \text{multiply 7 by 2} \\ = 14 - 6 + 3 & \text{add 6 and 3} \\ = 14 - 9 & \text{subtract 9 from 14} \\ = 5 & \end{array}$$

There may be other ways of changing the expression that lead to a result of 5. For example, she might have misread the + sign between 6 and 3 as a - sign. However, this misreading while possible is less likely than confusing the order of operations as illustrated above.

At this point, we have an idea of what happened on the first problem, but one example is not enough. We still can't choose from the range of reasons that could explain why she made the error. Is she "chunking" the expression and doing the left-hand part and the right-hand part separately then combining with the middle operation? Is she adding before subtracting? Is she inserting parentheses in ways that we don't yet understand?

In the second example, Alexis should have done the following:

$$\begin{array}{ll} = 9 - 5 + (16 \div 8) & \text{divide 16 by 8} \\ = 9 - 5 + 2 & \text{add 5 and 2} \\ = 9 - 7 & \text{subtract 7 from 9} \\ = 2 & \end{array}$$

However, she arrived at 2 as a result instead of 6. What might she have done that would explain this?

There is a pattern emerging: Alexis seems in both cases to have added then subtracted where she ought to have subtracted and then added. But again, we still do not know why she made the error. It seems less likely now that she might be chunking, or dividing the problem into left and right sides, since doing so would have yielded a correct answer on Example 2. She could be adding before subtracting. While there is evidence here that she uses parentheses correctly, we still don't know exactly why she divided first. Was it because of parentheses or because the operation was division?

In the third example, Alexis solved correctly:

$$\begin{array}{ll} 9 + 24 \div 3 - 1 & \text{divide 24 by 3} \\ = 9 + 8 - 1 & \text{add 9 and 8} \\ = 17 - 1 & \text{subtract 1 from 17} \\ = 16 & \end{array}$$

She also could have subtracted 1 from 8 then added 9 and 7 but, in this case, it happens to make no difference. It is also less likely, given her inclination to add before subtracting to this point.

What do we know from the third problem? The emerging pattern of addition before subtraction holds (here it does not happen to be incorrect to add before subtracting.) We also know that she divided before adding or subtracting, even in the absence of parentheses.

In the fourth example, Alexis solved correctly:

$$\begin{array}{ll} 17 - (3 + 7 \times 2) & \text{multiply 7 by 2} \\ = 17 - (3 + 14) & \text{add 3 and 14} \\ = 17 - 17 & \text{subtract 17 from 17} \\ = 0 & \end{array}$$

You can also use familiarity with common student misconceptions to help think about and answer this assessment task. Order of operations is often taught using the mnemonic PEMDAS, which is often referred to by a label such as “Please Excuse My Dear Aunt Sally.” This mnemonic is meant to indicate that one should complete, in order, parentheses, exponents, multiplication/division, and addition/subtraction. The use of PEMDAS is handy but can lead to a number of misconceptions. A common misconception is that the operations go strictly in order as listed—that is, that multiplication always comes before division and that addition always comes before subtraction. Since we know that Alexis is adding before subtracting, it is likely that this is the underlying cause of the errors. It was not necessary to know this information to solve the problem, but familiarity with this common error would make it easier to figure out what Alexis is doing.

### **What can you conclude about Alexis's thinking?**

- Alexis correctly does multiplication and division before either addition or subtraction.
- We have some evidence that Alexis correctly does work in the parentheses first, although this is not conclusive.
- Alexis incorrectly does addition before subtraction, except in cases where the subtraction is in parentheses.
- Alexis's work suggests she may be relying on an incorrect understanding of a mnemonic such as PEMDAS to solve the problems.
- There is no apparent pattern of incorrectly chunking the work.

### **What is the rationale for selecting an answer?**

At this point, we know the likely cause of the error (addition before subtraction) and must decide which of the options would be answered incorrectly if this pattern were continued for the next four problems.

#### **Option A:**

In Option A, addition before subtraction would give a correct answer, so Alexis is not likely to answer incorrectly.

#### **Option B:**

In Option B, if Alexis continues the same pattern of doing addition before subtraction she will answer incorrectly. Addition before subtraction is incorrect for this problem and there are no parentheses to guide her as there are in Option C. She will likely answer incorrectly.

#### **Option C:**

In Option C, if we assume that Alexis does the work in parentheses first, she would answer correctly.

#### **Option D:**

In Option D, once again, if Alexis correctly uses parentheses, she would get the right answer.

Option B is the one that Alexis is most likely to answer incorrectly, and it is the best option.

### **Summary of key knowledge, skills, and reasoning**

#### **This assessment task draws on:**

- Knowledge of how order of operations is used to evaluate numerical expressions.
- Familiarity with common student errors with order of operations, including confusing the order of operations and incorrect use of strategies such as chunking.

- Awareness that tools, methods and other aids that are commonly taught to help students solve problems can also support student misconceptions.
- Ability to identify problems that are likely to reveal a student error or source of confusion.
- Ability to analyze student work to identify the steps that were used to arrive at correct and incorrect solutions.

**Figure 1. Mathematics task design rationale.**

**A coding example: The Santiago item.** Each TDR represents a design hypothesis for what correct reasoning should look like but also represents a thorough and lengthy explanation of that reasoning, which is far more detailed and polished than would be reasonable to expect an interviewee to provide during a spoken interview. Some participant responses are simply more detailed than others, and one challenge was coding for whether the reasoning reflected the essence of our design hypothesis and not simply the level of detail provided. One coding task was to specify for each item the essential information that had to be present in the response for it to count as conforming, what we referred to as the *minimal conformity test*. These conformity tests assisted coders in maintaining consistency and also served to characterize the nature of a conforming response in cases where there might be multiple valid reasoning paths. For example, for the Santiago item, there were three possible conformity tests of a participant’s reasoning. To conform, the participant must

1. note that Alexis did addition before subtraction, based on evidence from the stimulus and that this would cause an incorrect answer for Option B (the key),
2. observe that Alexis added before subtracting because of the acronym PEMDAS and that this would cause an incorrect answer for Option B (the key), or
3. conclude Alexis added before subtracting without explicitly stating why and that this would cause an incorrect answer for Option B (the key).

For a response to conform, it was essential that the interviewee identify Alexis’s error (addition before subtraction) and connect this error to her likelihood of answering Option B (the key) incorrectly. The conformity tests allow for some variation in both the reasoning used to decide what Alexis’s error is and the degree to which that reasoning is expressed explicitly. Tests 1 and 2 characterize two different ways that the interviewee might decide that Alexis’s error was

addition before subtraction—one based on careful analysis of the item’s stimulus and the other based on recognition from prior experience with this common student error.

Test 3 characterizes cases where the interviewee correctly identified the error and connected this to the selected answer, but did not explain how Alexis’s error was identified. While such responses might be somewhat unsatisfying compared to cases where the explanation was more detailed, we judged that these had to be included as minimally conforming answers because our interview protocol did not prompt the interviewee to provide detail on this point. So this code was used to mark such cases as minimally conforming, provided that no other information in the response indicated problematic reasoning. Of 12 conforming responses to this item, six responses were coded on Test 1 only, indicating an explicit analysis of the stimulus information determined Alexis’s error, three were coded on Test 2 only, indicating that a reference was made to this known error type without mention of the item stimulus, and two were coded on both Tests 1 and 2, indicating that the interviewee provided both parts of the reasoning. Only one case was coded as conforming under Test 3.

An example taken from a conforming response is shown below. This response was coded as conforming with the TDR under Test 1, indicating that the interviewee reasoned from the given stimulus to identify Alexis’s difficulty.

Interviewer: Okay, if you could walk me through your thinking in arriving at Option 2 as your answer.

Respondent: Sure. The first step was to figure out what she was doing wrong so that I could figure out again which one she would probably compute wrong again.

At this point, the respondent has clearly identified the task of teaching, including both parts—the diagnosis of the given work and the prediction of how Alexis’s misconception will affect her work on the given problems.

And so I first solved it the correct way. So, for instance, in Number 1, you know I came up with 14 minus six plus three and then that’s eight plus three which is 11. And then I tried to figure out, how did she get down to five, and it did require quite some thought but I realized that she definitely did the 14, the seven times two, first. And so, in order to get... in order to go from 14 to five, she must have subtracted nine. So what did she do in order to get nine as her value? And so,



what I figured out was that she, in her mind, had actually computed the addition before the subtraction. And then that was so that when she added six plus three she got nine and then she took away nine instead of just taking away the six and then adding three.

The respondent, at this point, has analyzed the first stimulus problem exactly as outlined in the TDR. She has a theory of what Alexis is doing incorrectly and goes on to test this theory against the next stimulus problem.

So she was doing the addition before she was doing the subtraction and that was the same case in the second problem. She needed to do the parentheses first. She got two, but what she did was, again, instead of saying nine minus five is four and then adding the two, she actually did the addition first, so that she did nine minus five plus two which is seven. So that was her mistake. So she adds six before she subtracts and she doesn't understand that both of those operations are equal and you just do them from left to right.

This particular respondent does not explain her reasoning about the third and fourth stimulus problems, and the interviewer does not follow up. The respondent does refer to those problems later, so it is likely that she analyzed them but simply did not provide the information during the interview. The response is not, at this point, identical to the TDR, which represents a full and detailed explanation of *every* part of the reasoning, but the response is still coded as *conforming* to the TDR because the essential information about Alexis's error is clearly present and there is no other evidence to suggest any misunderstanding. The respondent goes on to the second part of the task.

So, in figuring out my answer, I tried to look and see where there was subtraction before the addition, because since her problem is that she was adding before subtracting, that's where she would have a problem is if the subtracting was actually coming first, and that's what she needed to do. So then I did the computation for each of the options that I looked at. In Option 2, the way it should be solved is to say three times two is six. So then you would have 13 minus six plus five and instead of going ahead and subtracting first—to say 13 minus six is seven and then five will give you 12—she was most likely to say 13 and then

instead of subtracting the six right away, she would add the six and five first to get 11. And so then she would do the subtracting next—which would be 13 minus 11, which is two. So her problem was that she was adding always before she subtracted when that wasn't following the order of operations.

Interviewer: Okay, so you said you did the computations for each option?

Respondent: Yes, I don't think I wrote them out, because I noticed that, for example, in the first one, the addition comes first, which is what she always does anyway, so she would still come up with the same thing. So if she were to solve it, she would say eight plus seven minus four when she does the 12 divided by three, and so that would—she would still get the same answer because she would add first and then do the subtraction. So she wouldn't come up with an incorrect answer. In Option 3 she still knows to do the parentheses first. She knows to do the division first, so 27 divided by three is nine and 'cause there are parentheses around the subtraction, she would still know to do the parentheses first because that was evidenced in Number 4 of her solution. So she understood to do—since she got that one correct—she knew to do the parentheses. And then also in Option 4, she knows to do the subtraction first, because it's in the parentheses. And so she knows to do four then times five plus 10 and there is no problem; it doesn't challenge her misconception of subtracting before adding or adding before subtracting.

The remainder of the response is a clear reflection of the reasoning in the TDR, outlining for each possible response how the identified misconception would or would not lead Alexis to answer incorrectly.

There were limited ways in which responses could conform to the item rationales/TDRs, but a wider variety of ways in which a response could fail to conform. These cases were also characterized by codes, including cases of explicit guessing or incorrect content knowledge or cases where the interviewee worked on a slightly different task of teaching. For example, one interviewee selected an incorrect answer on the basis of his teaching experience, stating that he generally found that students struggled with parenthesis, so he selected an answer option that featured parenthesis. The essential problem with this reasoning is that it does not characterize

Alexis's mistake, which was the task of teaching embedded in the item. This response is coded as "different task of teaching/answering a different question," as the response indicates that the interviewee is focused on anticipating student difficulties rather than diagnosing a particular student's error. In choosing to describe, instead, a student difficulty that he is familiar with, the respondent leaves Alexis out of the picture entirely. Another respondent answered similarly, based on what she expected might be difficult, summarizing at the end that "I didn't even solve the problems; I just looked at the kid's answers and I was just looking for an operation that might be difficult to do."

Others did understand the task of teaching but were simply unable to diagnose Alexis's error. One such respondent vacillated among the options, changing her mind several times and entertaining several theories of what Alexis might be misunderstanding, stating, at one point, "She just doesn't understand order of operations. So it's hard to say what she would really do, because she really doesn't understand the concept completely." Another incorrect response pattern was respondents who struggled to diagnose the error despite familiarity with the error type, a case that one might describe as difficulty in *applying* knowledge even if one has the appropriate knowledge. For example, the respondent below clearly understands the task of teaching and is frustrated that she cannot diagnose Alexis's problem. She also clearly understands how PEMDAS can lead to ordering multiplication and division incorrectly, a mistake quite similar to the ordering of addition and subtraction incorrectly, but knowing this and knowing what is to be done is not enough:

I just found confusing Number 2 on Alexis's test—that I could not figure out how she got two and I was very frustrated with that because, to me, if I don't know why my student shows an answer, I can't help them come up with the correct answer [and] the correct method to solve. But I figured out for Number 1 that she did minus three instead of plus three and that's how she got five, but I just could not figure out how she got two as her answer. I just made the assumption that maybe she doesn't really know her orders of operations and she just got lucky on Number 3 and 4 and she needs to be retaught and I really just thought that, [on] Option 4, she would not do her order of operations correctly as to where in Options 2 and 3, I felt like, and 1—they go kind of in the order. Like, it's more. How can I say this? It's more noticeable that you're going to do division first or

multiplication. I couldn't solve the second one but, I guess, like when I first glanced at it, I thought, "Oh well, it's probably going to be, you know, she used Please Excuse My Dear Aunt Sally," which we now know no longer works, because the students are strictly doing division before multiplication. They're not tying them together; whichever one you hit last first is the one you do. So we've had to go from Please Excuse My Dear Aunt Sally, which is a wonderful mnemonic device, to "No! You can no longer use that because you don't always do division first; you can do multiplication first, whatever you hit first left to right." So that's what I thought she would have done wrong but she really didn't.

Another respondent similarly explains the way in which PEMDAS can support misconceptions about ordering, including misordering addition and subtraction, but then concludes that Alexis worked backward, a conclusion that is not supported by the given stimulus problems.

Respondent: You know, typically, when we teach order of operations, you do have, you know, you'll have some children who will get questions wrong because they don't, because they don't know the order or they get confused about, you know, what order they should go in.

Interviewer: And can you tell me a little more about that?

Respondent: Just you know we teach them you know certainly to go parentheses, multiplication division, addition and subtraction and you know they take that literally to mean that I should go. They don't have much issue with the parentheses; they all know that you know I've got to solve everything that's in parentheses first but then from there they tend you know they you know they may because it's multiplication, division, addition, subtraction. They may go in that particular order versus what it really means is that you do all of the multiplication and all of the division, then all of the addition and all of the subtraction. But it doesn't mean that you have to go in that order. So if you see multiplication, you see division first, then you do the division, and then the multiplication. If you see subtraction after that, then you do subtraction rather than having it be that "I got to do all the multiplication, then all of the division then all of the addition then all

of the subtraction.” So, you know, just the procedure that we teach sometimes throws the kids off.

Interviewer: All right. Thank you for that and, now, you had selected Option 1 as your answer. Can you tell me as to why you selected this as your best answer?

Respondent: My notes here say it looks like the answers that she got incorrect she worked—were based on her working—backwards and so that, to me, seemed the most logical one, where working backwards would give you the wrong answer.

Another pattern of incorrect reasoning was attending to surface similarities between the stimulus problems and the options rather than considering *why* Alexis answers incorrectly. One respondent eventually chose Option 1 because “it had all of the same operations as Number 2. So I was trying to find something that had the same operations that the student might struggle with.”

**The importance of training with data.** In the section above, we presented an example from the Santiago item. We have attempted to capture in print the kind of details we considered in our coding discussions and how the team used evidence to learn how to code items. One reason that cognitive interview studies of this sort are time-consuming is that learning to code requires significant investment in working with actual data and using the data to clarify shared understanding of coding rules for agreement. In this study, the TDRs played a critical role in specifying the expected reasoning and thus study hypothesis for agreement. However, to ensure accurate coding, the research team needed to go through shared work for each item, discussing the TDR and selected data responses, and reaching consensus around criteria for conformity to the TDR.

## **Results and Discussion**

### **Results for Research Question 1**

The first research question asked, “To what extent do teachers’ classifications on assessment score (MET) correspond to their classifications on conformity with the intended reasoning?” We addressed this research question by looking at the distribution of conformity of responses over the lower and higher performing quartile groups. This analysis technique was only moderately successful. Results are presented first, and a discussion of limitations of this method follows. The unit of analysis for Tables 5 and 6 are item responses, which is to say that each individual’s response to each item is considered as a separate case.

For mathematics, responses from interviewees with lower MET scores were less likely to conform with the TDR, and responses from interviewees with higher MET scores were more likely to conform. Group differences are highly significant ( $p < .01$ ) and the association between MET quartile classification and conforming justification is strong ( $\chi^2 (1, N = 640) = 80.97$ ).

**Table 5**

*Conformity of Mathematics Responses to TDR by Quartile Group*

	Lower performing quartile	Higher performing quartile	Total
Justification conforms to TDR	141	244	385
Justification does not conform to TDR	186	69	255
Total	327	313	640

*Note.* TDR = task design rationale.

For ELA, the associations between quartile grouping and conforming of reasoning are weak ( $\chi^2 (1, N = 540) = 3.91$ ) but still statistically significant ( $p < .05$ ).

**Table 6**

*Conformity of ELA Responses to TDR by Quartile Group*

	Quartile 2 group	Quartile 4 group	Total
Justification conforms to TDR	164	185	349
Justification does not conform to TDR	106	85	191
Total	270	270	540

*Note.* ELA = English language arts; TDR = task design rationale.

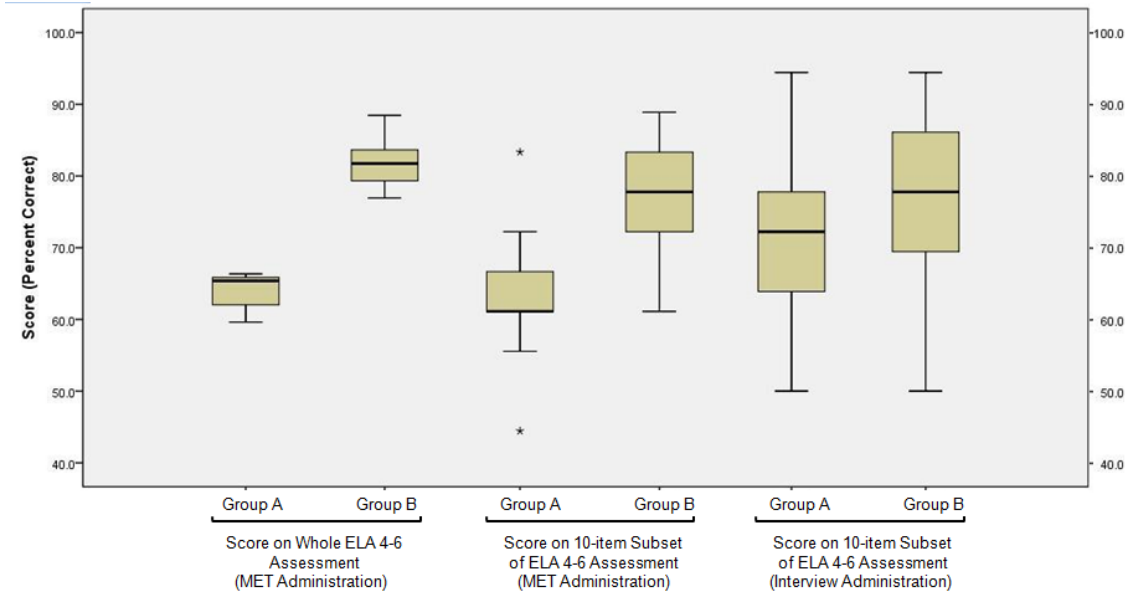
One of the limitations of the quartile group analysis is that we are extrapolating from the subset of items included in the cognitive interview to the overall score on the full assessments administered in the MET study. The inference that participant reasoning extrapolates to this score depends on assumptions that the interview subset score is representative of the full assessment score and that that the scores generated in the test-taking environment are comparable to those generated in the cognitive-interview environment. There are two specific ways in which these assumptions may be suspect.

First, the subset of items used in the study interview may not be representative of the full assessment administered in MET. Recall that the items were purposefully selected from the

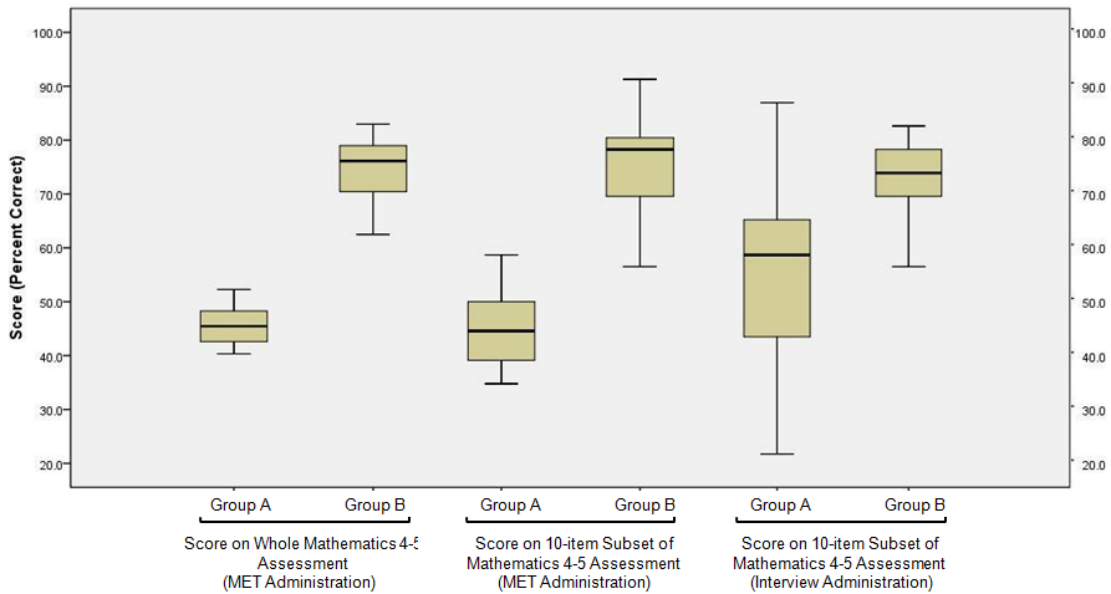
overall assessment to represent a number of features of interest. This selection process was not random and therefore it is highly likely that the subset of items is not representative of the full assessment. To better understand the strength of the inference we are making to answer this research question, we examined the extent to which quartile classification based on responses on the subset of items in the interview correspond to quartile classification based on the full assessment. For the entire MET sample, 59% of the mathematics teachers and 56% of the ELA teachers originally in Quartiles 2 and 4 remain in their respective quartile when only the 10-item subset is considered. Among the teachers in our interview sample, 61% of mathematics teachers and 53% of ELA teachers remained in their original quartile. This suggests that the subset of items selected for the cognitive interview may be measuring something different than the full set of items. Some movement from one quartile to another is to be expected when using only 10 items to score teachers, as tests that are this short cannot achieve high levels of reliability. We expect that any subset of 10 items would lead to some changes in how teachers are rank ordered. In addition to the decreased reliabilities we would expect for any 10-item subset, these differences in ranking could reflect a number of things, including differences in overall difficulty between the original assessments and the 10-item subsets. We do not here explore additional analyses that might clarify why we observed these differences, but point out that these differences suggest limitations in the strength of the inferences that can be drawn from the 10-item subset to the full test. If extrapolation from a subset of items to an assessment score is desirable and one wishes to use performance groupings based on assessment score, items should be carefully selected to support that extrapolation by preserving those performance groupings to the greatest extent possible.

Second, we noted in conducting the interviews that the answers given by participants for the 10-item set during the MET administration and the answers given during their interviews on those same 10 items differed. In both mathematics and ELA, the mean scores for participants originally in Quartile 2 on the 10-item subset at the MET administration increased slightly, from 63% to 69% in ELA and from 45% to 56% in mathematics. In ELA, for both quartile groups the range of scores widened, indicating that between original administration and interview administration some teachers were more likely to answer items correctly and some more likely to answer incorrectly, regardless of their original quartile status. In mathematics, the same pattern was evident for participants originally in Quartile 2 but those originally in Quartile 4 were likely

to maintain high scores. The differences in group ranking for Group A (Quartile 2) and Group B (Quartile 4) for the score on all MET items, on just the 10 items, and on the answers given in the cognitive interviews are presented for ELA in Figure 2 and for mathematics in Figure 3.



**Figure 2. Distribution of interview participants’ English language arts scores.**



**Figure 3. Distribution of interview participants’ mathematics scores.**



The observed differences could be due to many reasons. We noticed that the amount of change in answer varied significantly by item, and item selection may have influenced the degree to which answers changed. A significant amount of time (about a year) passed between MET administration and the interviews, and participants may have learned (or forgotten) during that time. There also seems to be some relationship between item difficulty and likeliness of answer to change, with fewer answer changes on items that were much easier, but the relationship is not sufficient to explain all of the change and, in some cases, items with similar levels of difficulty elicited different levels of answer change. Another possibility is that the context of the cognitive interview study led participants to approach the test differently. One might expect that knowing they would be interviewed would lead participants to spend more time and care in answering in preparation for the interview. It is also possible that the demand of explaining why nonselected options were discarded as well as why the selected option was chosen might prompt a more thorough approach to the item at the time of the interview. This explains ELA results less well than those in mathematics, however, because one would expect a more careful approach—either in advance or during the interview—to lead only to improved performance. A final possibility that there is simple a degree of random error that may cause variation in answers at different administrations of the same test.

This suggests that, for both ELA and mathematics, the selected items used in the cognitive interviews rank order teachers differently than do the overall assessment scores. It is particularly troubling that in ELA, score distributions for the two groups on the cognitive interview scores suggest that, at least for the subset of items in the interviews, the two groups may not really represent distinct performance groups, making their comparison as such less meaningful. While this doesn't invalidate the warrant for the claim that scores on the CKT assessments represent the knowledge and reasoning we are seeking to measure, with strong evidence in mathematics and moderate evidence in ELA, it does suggest that this is not a very strong warrant, given the classification alignment between assessment items and cognitive interview items. It also suggests that, especially for ELA, comparison by performance groups may be questionable as a result. These results suggest, too, that the processes used to answer items during the MET testing administration may differ from those used in the cognitive interview administration, calling into question the methodological assumption that the reasoning

used in cognitive interviews can be directly extrapolated to the reasoning used in the regular course of testing.

## Results for Research Question 2

Research Question 2 asked, “For each item and for all items, to what extent is correct and incorrect reasoning associated with correct and incorrect answers, respectively?”

This research question addresses the alignment of reasoning with correct/incorrect answers, supporting claims that our items are correctly keyed and that the reasoning employed in reaching correct answers approximates the use of knowledge and skills embedded in the TDRs. This question differs from the first research question in its focus on just the set of items used in the interviews. We cannot make strong validity claims for the MET assessment forms based on only this subset of items, particularly in the context of the limitations discussed above. However, the results shown in Tables 7 and 8 provide strong evidence that the reasoning that these items capture is the reasoning that the items are designed to capture and that the items are correctly keyed such that correct or incorrect answers give a very good signal of conforming or nonconforming reasoning.

For mathematics, 88% of responses fall on the diagonal, indicating that an incorrect answer reflects nonconforming reasoning and a correct answer indicates conforming reasoning. For ELA, almost 90% of results fall on the diagonal. As in the previous analysis, the unit of analysis is the item response.

**Table 7**

*Alignment Between Correct/Incorrect Answer and Conforming Reasoning for Mathematics*

	Response does not conform to TDR	Response conforms to TDR	Total
Answered incorrectly	<b>185<sup>a</sup></b>	6	191
Answered correctly	70	<b>379<sup>a</sup></b>	449
Total	255	385	640

*Note.* TDR = task design rationale.

<sup>a</sup>In this table, 88.1% (564) of total responses (640; in **bold**) fall along the diagonal; this indicates that correct answers reflect conforming reasoning and incorrect answers reflect nonconforming reasoning, as measured against the TDRs.

**Table 8**

***Alignment Between Correct/Incorrect Answer and Conforming Reasoning for ELA***

	Response does not conform to TDR	Response conforms to TDR	Total
Answered incorrectly	<b>141<sup>a</sup></b>	4	145
Answered correctly	50	<b>345<sup>a</sup></b>	395
Total	191	349	540

*Note.* ELA = English language arts; TDR = task design rationale.

<sup>a</sup>In this table, 90.0% (486) of total responses (540; in **bold**) fall along the diagonal; this indicates that correct answers reflect conforming reasoning and incorrect answers reflect nonconforming reasoning, as measured against the TDRs.

Because conforming to the TDR implies a correct answer was reached, it was not possible to code an item as conforming if the respondent gave an incorrect answer. However, a small number of items (six for mathematics and four for ELA) do appear in the tables above with this coding. Most of these were cases in which the respondent initially gave an incorrect answer but then changed to a correct answer during the justification of the answer. These cases could reasonably be recoded as conforming, which would strengthen the listed results. One was a case in which the respondent answered ambiguously such that it was not possible to determine in analysis what answer had been selected, although the justification offered conformed to the TDR.

Most of the off-diagonal responses representing nonconforming reasoning leading to a correct answer occurred on table items. Removing table items, 93% of mathematics and ELA responses fall on the diagonals, indicating a high degree of alignment between item answer and use of intended reasoning and knowledge. These off-diagonal cases are further explored in the third research question.

**Results for Research Question 3**

Research Question 3 asked, “For responses for which correct/incorrect reasoning does not associate with correct/incorrect answers (at both the item level and for all items), to what extent and in what ways are these responses due to:

- a. defensible reasoning that supports a different item key?
- b. nondefensible reasoning that is associated with the correct key?”

**Defensible reasoning that supports a different item key.** There are two ways that defensible reasoning supporting a different key might have been visible in our results. The first,

item responses coded as incorrect answers but conforming reasoning, was discussed above. The majority of such cases were cases in which a respondent changed from an incorrect to a correct answer. Such cases might represent construct-irrelevant variance in the overall score, but there are few of them, and they are not a form of systematic construct-irrelevant variance and cannot be controlled for through item design.

The other way that defensible reasoning supports a different key is in the application of the global code “defensible non-TDR reasoning,” which coders used to flag such cases. Overall, only 32 item responses were coded as defensible non-TDR reasoning somewhere in the response, about 2.7% of all item responses. In most cases where the code was applied, it was applied only to a portion of the item response and most often represented an alternate way of rejecting one of the distractor choices rather than a justification for a different key. These codes point out to us places that the TDRs may require minor revision to account for alternate but valid ways of rejecting the distractors, but they do not indicate incorrect keying of the item.

**Nondefensible reasoning that is associated with the correct key.** In mathematics, 70 item responses (11%) were coded as correct answers with reasoning that did not conform to the TDR. In ELA, there were 50 such cases (9.5%). In mathematics, only three of the 70 were coded as explicit guessing/uncertainty, which does not explain the majority of cases. However, 58 such cases are on table items and only 12 on nontable items. Similarly for ELA, only three of the 50 are coded as explicit guessing/uncertainty, but 35 of these are on table items and only 16 on nontable items. MC items and table items are different from one another in several ways that might explain this difference. Table items present the test taker with only two answer options, increasing the chances of guessing correctly, and it is possible that these cases were the result of guessing or informed guessing even if the interviewee did not explicitly report having guessed. Another type of difference has to do with how item responses were coded. Many responses to table items were shorter and less complete than those given to MC items, possibly because responding to a set of items in the context of one larger question inclined the participants to say less about each individual piece or to assume that statements mentioned with respect to an earlier item in the question carried over to discussion of a later item in the same question. As a result, many responses were coded as nonconforming due to the explanation being insufficient for coders to draw a conclusion about conformity to the TDR, not because there was evidence of incorrect or nonconforming reasoning. It is possible that these cases of correct answers with

nonconforming reasoning might simply reflect correct answers with correct reasoning that was explained in insufficient detail.

#### **Results for Research Question 4**

Research Question 4 asked, “To what extent do the items remind teachers of something they have experienced in their teaching? To what extent do teachers perceive the items to be authentic problems that would be encountered in teaching?”

We found that 97% of responses in mathematics and 96% in ELA indicate that the items remind the interviewees of something they have experienced in their teaching or that they are problems they expect other teachers might encounter even if they have not. This speaks to the face validity of the items, and strong face validity can reduce construct-irrelevant variance that may result from test takers feeling that test items are irrelevant to them.

#### **Summary and Conclusion**

The primary goal of this study was to validate the scoring and extrapolation inferences for CKT assessments designed for teachers of Grades 4-5 mathematics and Grades 4-6 ELA. We found strong evidence to support the scoring inferences for these items. The inference that CKT reasoning as demonstrated in interviews extrapolates to overall score performance was moderate for mathematics and weak for ELA. We noted that the scores for the cognitive interview teacher participants differed across three scoring conditions: whole assessment, 10 item MET subset, and cognitive interview responses for the 10 item subtest. These differences suggest that our extrapolation inference is influenced by item selection and study context. We found much stronger evidence, for both ELA and mathematics, that the reasoning used by participants to answer the items conformed to our item-design theory, suggesting that the items are eliciting the desired CKT knowledge and extrapolating to the target domain. Because this study was conducted with only two assessments and teachers from just a single district, we have no evidence that the validity claims extend beyond this limited sample. However, the results do suggest that this type of CKT item design is valid across ELA and mathematics, two very different subjects. This supports an initial hypothesis that the design may also be valid across other subject differences.

We also presented a detailed account and associated explanation of the methods used to conduct the study. Our goal was to share the details of this study as an example of a particular

methodology and thus inform the reader of its benefits, drawbacks, and considerations that need to be taken into account. Due to time constraints, cognitive interviews are generally conducted on a relatively small number of items, and as our analyses demonstrated, this can create problems in extrapolating results to overall performance. This problem may be due in part to how much item-selection criteria attempt to preserve alignment between item performance and overall performance but are also a reflection of the lack of reliability any sufficiently small subset of items will have and, therefore, an inherent limitation of the methodology. It also suggests that there may be ways in which the situation of a cognitive interview influences participant reasoning such that the connection between test performance and interview performance becomes tenuous. Design considerations for such a study include the tradeoffs between concurrent and retrospective think-aloud approaches, which are in part constrained by the conditions of the study but also depend on the research questions being addressed and what types of reflection are most likely to serve in answering them. Similarly, decisions about conducting interviews in person or by phone, about how far in advance to provide the instruments, and about how much participant training can be built in depend heavily on study context, but the effects of those choices also depend on study context and may affect results more or less in different studies. Because the number of interviews possible is likely to be small, participant selection is an important consideration and requires attention as to whether one wants a representative sample and on what characteristics it should be representative. Interviewer training is crucial, but the degree to which background knowledge and content-specific training is necessary in addition to general training will vary by study and depend on the nature of the claims that one desires to make. In coding data, decisions about using deductive or inductive coding or a combined approach depend directly on the research questions, and we suggest that when using a validity framework approach, it may make sense to use documents similar to our TDRs to specify the interpretive argument at the item level. Because the item is the unit of analysis that interviews lend themselves to.

We also recognize that there are ways of analyzing this data set that we have yet to explore, including aggregation at the item level (to help us learn more about the functioning of particular items) and at the individual level (to help us learn more about trends in a teacher's approach to such items, including trends that might appear across the content areas of mathematics and ELA). These analyses could augment the validity evidence presented in this

report by providing more nuanced information on successful and less successful item-design characteristics and by illuminating those aspects of teacher knowledge and reasoning that are specific to an item, those that are more generic approaches that may fall into different classes of how individual teachers reason, or those that may indicate how the knowledge and reasoning is subject-specific or more generically applied across subject.

We also see this study as a promising approach to Kane's (2006) extrapolation inference, particularly during the development phase of a new measure. Extrapolation is often evaluated through comparison to criterion measures, but this approach is limited by the types and quality of the criterion measures available. In the case of this study, there are no alternate measures of CKT with more strongly established validity that might be used as comparison points for our measures. In such cases, one can use measures of related constructs, but the utility of this comparison will be limited by the extent to which the constructs are related. For example, one might examine relationships between CKT scores and scores on traditional content measures on the assumption that strong content knowledge is a necessary condition for strong CKT, but because it is a necessary but not sufficient condition, these correlations should be expected to be weak at best. Alternately, one might examine the relationship between CKT measures and measures of instructional quality (as the recent MET study did) but again, the expected relationship would be weak because CKT is a necessary but not sufficient condition for strong content instruction. These relationships would also depend greatly on the extent to which the instructional quality measures focus on content instruction. This is not to say that such comparisons are not useful, but interpretation of the results will require caution and will be difficult to do without accompanying qualitative data to help in interpreting why there are or are not relationships. In such a situation, methods such as those described in this report, despite their inherent limitations, can provide more direct validity evidence to support claims that the assessment measures the construct of interest.

## References

- Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In R. Lesh & A. Kelly (Eds.), *Handbook of research design in mathematics and science education* (pp. 547–589). Hillsdale, NJ: Lawrence Erlbaum.
- Ericsson, K. A., & Simon, H. A. (1985). *Protocol analysis*. Cambridge, MA: MIT.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.
- Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology*, *113*(3), 387–404.
- SocioCultural Research Consultants. (2012). *Dedoose* [Computer software]. Retrieved from <http://app.dedoose.com/>



## Appendix A

### Task Design Rationales for Mathematics Items

#### Task Design Rationale – Jimenez: Nature of the Remainder

In a unit on division, Ms. Jimenez gave her students the following problem.

Carlos wants to cover the bottom edge of his window with a row of tiles that are each 5 inches long. If the bottom of the window is 42 inches long, how many tiles will he need to buy?

When the students finished, she wanted to give them another problem that was mathematically similar to the first. Of the following, which is most similar to the original problem, both in terms of the meaning of division and the nature of the remainder?

- Tim needs to pack 42 binders into boxes in order to ship them. If each box can hold 5 binders, what would be the fewest number of boxes needed to ship all of the binders?
- Gabriela has 42 stickers and wants to divide them up equally among 5 of her friends. How many stickers should each friend get?
- Robert wants to wrap gifts, and he has 42 feet of ribbon on a spool. If each gift requires 5 feet of ribbon, how many gifts can be wrapped using the ribbon on the spool?
- Teresa has 42 inches of string. If she cuts the string into 5 equal pieces, how long will each piece be?

#### What is this assessment task asking?

This assessment task asks you to identify a problem that is similar to a given one in terms of the meaning of division and the nature of the remainder. It turns out that knowing the nature of the division is not enough to complete the task, and you really have to consider what the remainder means in each problem, which you may do explicitly or may simply recognize as a pattern of similarity among the problems. It is also possible to answer correctly just by looking at the numeric answers to each problem in the options (9, 8, 8, 8, 4) and seeing which one matches the given problem's answer of 9. But the intention is to think about the meaning of the remainder, not just to see which problem has a matching answer.

#### What information is important?

The assessment task instructs you to consider the meaning of division and the nature of the remainder. In the given problem, and in each of the options, the divided values are identical:  $42 \div 5 = 8 \text{ remainder } 2$ . Carlos has a window edge to cover that is 42 inches long, and he has 5-inch tiles with which to cover it. The window has to be fully covered (you need enough tiles to cover it, even if that means you end up with leftover tile.) This means you need a whole number of tiles (you cannot buy part of a tile) and will have to round up, for an answer of 9 tiles needed. The essential attributes of the given problem are the following:

- The division problem represents what is called a *quotative* or measurement model. In other words, there is a given quantity (the length of the windowsill), and the problem asks you to measure that quantity using another quantity as

the unit of measure. It is as if you were measuring the windowsill using tiles as a ruler. This is different from the other common model of multiplication, the *partitive* model, which represents fair sharing by distributing equally among a number of groups.

- The answer represents the minimum number of the measuring quantities to completely cover the quantity being measured, which requires rounding up to the next whole tile. This is unlike a problem that asks how many of something will *fit* (that indicates rounding down) or that allows for partial units (decimals).

What is the rationale for selecting an answer?

**Option A:**

In Option A, Tim has 42 binders to ship in boxes that hold 5 binders each. In order to ship them all, he will need 9 boxes. The first 8 boxes will be full, while the remaining 2 binders will need to go in a separate box. This, like the given problem, is a quotative division model (you could think of it as measuring out the binders in five-binder groups). Also, like the given problem, the scenario indicates that all the binders must be shipped, and this means an extra box will be needed, even if it is shipped partially empty, since there are no partial boxes. So you must round up to the extra box. This problem is similar to the given one with respect to both the meaning of division and the nature of the remainder, and it is the best answer.

**Option B:**

Option B involves Gabriella dividing her stickers evenly among her friends. In this situation, each friend will receive 8 stickers, and it is unclear what happens to the remaining 2 stickers. Perhaps she keeps them, or perhaps two of her friends get 9 stickers each, even though the others get only 8. This problem is an example of the partitive division model, and neither of the possible outcomes indicates that rounding up would be appropriate (Gabriella does not conveniently happen upon three more stickers so that she can give each friend 9). This problem is not mathematically similar to the given one in terms of the meaning of division or the nature of the remainder.

**Option C:**

In Option C, gifts will be wrapped with 5 feet of ribbon each. Since there is only 42 feet of ribbon, only 8 gifts can be wrapped. This, like the given problem, is a measurement problem in which the 42 feet are being measured out in 5-foot lengths. However, this problem asks how many gifts can be wrapped, indicating that once the ribbon is used up, one will stop wrapping. It is asking how many 5s “fit” fully into 42, which indicates that rounding down is appropriate, for an answer of 8. This problem is similar to the given one in terms of the meaning of the division, but differs in the nature of the remainder.

**Option D:**

Option D involves cutting a 42-inch length of string into 5 equal pieces. Since there is no reason the 5 equal pieces must be whole number lengths, this problem permits an exact value answer of 8.4. Like Option B, this option represents a partitive division model, as it represents a fair division of the string into 5 pieces. There is no remainder. Neither the meaning of division nor the treatment of the remainder is mathematically similar to those in the given problem.

Option A is the best answer, as it is similar to the given problem in terms of both the meaning of division and the nature of the remainder.

Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Knowledge that division can be thought about in different ways, including a measurement (quotative) model and a fair-sharing (partitive) model; how each model works; and how the models are different from one another.
- Knowledge that there are different ways to think about the remainder of a division in a word problem, including rounding down, rounding up, or incorporating the remainder into the answer as a decimal.
- Ability to analyze the context of a problem and recognize its attributes in order to find similar or dissimilar problems.

## Task Design Rationale – Taylor: Associative and Commutative Properties

A lesson in Ms. Taylor's textbook states the associative and commutative properties of addition. To motivate the students to learn the properties, she tells her students that the properties can often be used to simplify the evaluation of expressions.

She wants to give her students an example that will focus their attention on how these properties can be useful in evaluating expressions. Of the following expressions, which would best serve her purpose?

- ( 455 + 456 ) + ( 457 + 458 )
- ( 647 + 373 ) + ( 227 + 456 )
- ( 551 + 775 ) + ( 49 + 225 )
- Each of these expressions would serve her purpose equally well.

### What is this assessment task asking?

This assessment task asks you to find the best example for a very specific purpose—motivating students to learn the associative and commutative properties. In particular, it directs you to choose an example that will focus their attention on the *usefulness* of the properties in evaluating expressions. This means you'll need to find the problem in which application of the properties makes the computation noticeably easier.

### What information is important?

It is important to know what each of the properties allows you to do in evaluating an expression. The associative property allows you to group numbers any way you like as long as they all use the same operation of addition or multiplication. In these examples, the operation is always addition, so this means that you can remove the parentheses and/or add new parentheses in order to add the values in different orders. For example, the associative property would allow you to regroup if you wanted to sum the middle two values first in Option A:

$$455 + (456 + 457) + 458.$$

The commutative property allows you to reverse the relative positions of two numbers, so if you wished to change the ordering of the last two values in Option A, you could do so:

$$(455 + 456) + (458 + 459).$$

Used together, these two properties allow you to add the four values in each problem *in any order you like*. For example, if you wanted to sum the first and last terms in Option A, you could rearrange as follows:

$$(455 + 456) + (457 + 458) = (455 + 456) + (458 + 457) \text{ by commuting 457 and 458}$$

$$\begin{aligned}
&= 455 + (456 + 458) + 457 \text{ by regrouping (associative property)} \\
&= 455 + (458 + 456) + 457 \text{ by commuting 456 and 458} \\
&= (455 + 458) + 456 + 457 \text{ by regrouping (associative property)}.
\end{aligned}$$

Since this problem is asking you to choose an example that illustrates how the properties are useful, this means you'll want to pick the example that is most noticeably easier to calculate by adding the values in a different order from that which is given.

What is the rationale for selecting an answer?

**Option A:**

The problem shown in Option A involves adding four consecutive three-digit integers, two at a time. There are no rearrangements that lead to obvious conveniences in terms of the place value—for example, the values in the ones places are 5, 6, 7, and 8, none of which pair nicely to give a result of 10. Because these are consecutive numbers, it is possible to simplify computation slightly if you know that the sum of the first and fourth numbers will give the same value as the sum of the middle two:

$$\begin{aligned}
(455 + 456) + (457 + 458) &= (455 + 458) + (456 + 457) \\
&= 913 + 913
\end{aligned}$$

This simplifies the arithmetic a little, if you know the trick, since you don't have to actually add 456 and 457 once you've added 455 and 458 (because they will also sum to 913), but there is still a fair amount of computation involved. This problem is made a little easier by application of the properties but only if you are familiar with the pattern that sums of consecutive numbers demonstrate and even then it is only a little easier. Option A could serve Ms. Taylor's purpose, but does not seem like an ideal example to show students how useful the properties can be.

**Option B:**

The problem shown in Option B is  $(647 + 373) + (227 + 456)$ . A student might notice that 373 and 227 sum to 600 and rearrange as  $(227 + 373) + (647 + 456)$  so that these numbers are together. However, the second computation is not made easier in this case, so this appears to simplify things only partially. This problem would do a better job of serving Ms. Taylor's purpose than Option A, since it doesn't depend on knowing a particular trick, but an example in which both parts were made simpler by rearranging would do a better job of illustrating the usefulness of the properties.

**Option C:**

The problem in Option C is one that becomes much simpler by rearrangement of the numbers so that they can add to powers of 100 in the first step:  $(551 + 49) + (775 + 225) = 600 + 1000$ .

This is a much easier computation than adding them in the given order, so this would highlight very effectively how the properties can be used to make computation easier. It serves Ms. Taylor's purpose well.

**Option D:**

Since Option C is better than A or B, the options would not each serve Ms. Taylor's purpose equally well.

Option C, which is rendered dramatically easier by rearrangement of the addition, best serves Ms. Taylor's purpose of illustrating to her students why the properties are useful.

Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Familiarity with the associative and commutative properties and an understanding of how they are applied, particularly in situations in which there are many values in an expression.
- Recognition that in combination, the two properties allow the addition to be performed in any desired order.
- Understanding that a compelling example of the usefulness of the properties will be one that is much easier to calculate once the properties are applied.

## Task Design Rationale – Xavier: Representations of Division

Mr. Xavier is teaching his students to solve word problems. Students are learning to draw pictures to help them reason about the situations and about the computations used to solve the problems. Mr. Xavier wants to give his students problems where each problem has a figure and a computation and have his students discuss in small groups whether the figure can be used to model the given computation. For each of the following, indicate whether the figure **can** or **cannot** be used to model the given computation.

Figure	Computation	Can Be Used	Cannot Be Used
	$12 \div 3$		
	$1\frac{1}{2} \div \frac{1}{6}$		
	$48 \div \frac{5}{8}$		
	$1\frac{1}{2} \div \frac{1}{3}$		

### What is this assessment task asking?

This assessment task is asking about the connection between a representation and the computation it is supposed to illustrate. It is not asking you if the representations are good choices in the sense of illustrating clearly or are ones that students would understand easily, just whether they could, in fact, be used. In general, a representation could be used if it makes sense and actually represents the underlying mathematics correctly, so this is what you'll need to pay attention to. Essentially, this assessment task comes down to figuring out the following:

- Can the figure be interpreted to be asking a meaningful question to which the “?” is the answer?
- Is this the same question posed by the computation?
- Is the answer to each question the same?

You may recognize a common student error in Row C, and noticing this makes it easier to respond to this row, but it is not necessary for you to know this to answer the assessment task.

### What information is important?

You'll need to consider each row independently, keeping in mind that the goal is to compare the question asked in each of the two pieces (the representation and the computation).

What is the rationale for selecting an answer?

**Row A:**

The figure in Row A is a bar that is 12 units long and has been split into three (seemingly) equal parts. Since there is a question mark below one of the parts, it is implied that the question posed in the figure is “What is the size of each part?” One of the ways of interpreting the computation  $12 \div 3$  is that it asks how big each group will be if 12 is divided evenly into three groups, which is the same question.

The result of the computation also matches the value of the question mark in the figure:  $12 \div 3 = 4$ , and in the figure, if the question mark were 4, three of them would give the indicated length of 12 for the entire rectangle.

This figure can be used to model the given computation.

**Row B:**

The figure in Row B is a bar that is  $1\frac{1}{2}$  units long and is split into pieces that are each  $\frac{1}{6}$  of a unit long. The question mark in the last box is a bit ambiguous, but taken together with the 1 and the 2 (and taking into consideration that the size of each piece is already specified as  $\frac{1}{6}$ ), it implies counting the boxes to determine how many such  $\frac{1}{6}$ -unit pieces fit into the  $1\frac{1}{2}$  units. One interpretation of the computation  $1\frac{1}{2} \div \frac{1}{6}$  is that it asks, “How many groups of size  $\frac{1}{6}$  will fit into  $1\frac{1}{2}$ ?” which is the same question.

The result of the computation also matches the value of the question mark in the figure  $1\frac{1}{2} \div \frac{1}{6} = 9$ , and in the figure, nine boxes are shown.

This figure can be used to model the given computation.

**Row C:**

The figure in Row C is a bar that is 48 units long and is split into eight parts. The question mark refers to five of those parts, effectively asking how much of 48 one has if five of eight equal pieces are taken. This question is equivalent to the multiplication of  $48 \times \frac{5}{8}$ . Since Row C models the multiplication of 48 and  $\frac{5}{8}$ , it cannot also model division of 48 by  $\frac{5}{8}$ .

The result of the computation also gives a different value from that illustrated by the figure:  $48 \div \frac{5}{8} = 76.8$ . Using the figure, 48 divided into eight equal pieces means that each piece is size 6, and five such pieces would be size 30.

This figure cannot be used to model the given computation.



**Row D:**

The figure in Row D is a bar with unknown length split into three equal pieces that are each  $1\frac{1}{2}$  units long. The figure seems to pose the question, “What is 3 times  $1\frac{1}{2}$ ?” While the given computation is division rather than multiplication, division by  $\frac{1}{3}$  is mathematically equivalent to multiplication by 3.

The result of the computation also matches the value of the question mark in the figure  $1\frac{1}{2} \times 3 = 4\frac{1}{2}$ , and adding up the three  $1\frac{1}{2}$ -sized boxes in the figure would give a total length of  $4\frac{1}{2}$ .

This figure can also be interpreted as a representation of a quotative model of division of  $1\frac{1}{2}$  by  $\frac{1}{3}$ .

While the interpretation of this one is perhaps less obvious, the figure can be used to model the computation.

Rows A, B, and D can be used to model the given computations; Row C cannot.

**Summary of key knowledge, skills, and reasoning**

This assessment task draws on the following:

- Awareness that division can be thought of in a number of different ways, including the quotative model (How many times does one quantity fit into another?), the partitive model (How big will each group be if the whole is divided into a certain number of equal groups?), and as the inverse of multiplication.
- Awareness that representation of a computation can be used only if it makes sense and is mathematically consistent with the computation.
- Awareness that confusing the representations of division and multiplication is a common student error, particularly in the case of dividing by a fraction.

## Task Design Rationale – Chamberlain: Meaning of Equals Sign

Mr. Chamberlain is concerned that his students' use of the calculator has led them to view the equal sign as a signal to carry out an operation rather than as a symbol indicating equality. Of the following missing-number problems, which would best assess whether students understand the mathematically correct meaning of the equal sign?

$\_ + \_ = 18$

$7 + 5 = \_ + 6$

$\_ = 17 + 9 + 5$

$23 + 4 = \_ = 4 + 23$

What is this assessment task asking?

This task asks you to determine which problem best assesses whether students understand the equals sign as an indicator to carry out an operation or as a symbol indicating equality. To complete this task, you first need to understand how the equals sign can be misunderstood by students as a command to carry out an operation and how this could lead students to interpret missing-number problems. Then you need to consider, for each problem, how a student with this misunderstanding might think through and answer each of the missing-number problems and which missing-number problem is most likely to lead to an incorrect answer.

What information is important?

It is necessary to recognize what it means to use the equals sign as a command to carry out an operation. One common way students develop this misunderstanding is through their work with calculators. Students can conclude that equals is a command to “do the math” for some given expression. For example, for the expression  $2 + 4$ , a student with this misunderstanding would think that pressing the equals sign on the calculator does the addition work and gives you 6 as the answer. A mathematically correct meaning of the equals sign is that  $2 + 4$  and 6 have equal values.

What is the rationale for selecting an answer?

### **Option A:**

This option involves a single number on one side of the equals sign. This would likely reinforce the idea that the value 18 is equal to the result of performing the operation on the opposite side of the equals sign. Students who understand the equals sign as a command to carry out the operation could still get a correct answer for this problem, and for this reason it would not be a good choice to assess whether students understand the mathematically correct meaning of the equals sign.

**Option B:**

In Option B, both sides of the equals sign are expressions. A student who understands the correct meaning of the equals sign could go through a reasoning process like the following: “I know that  $7 + 5$  is 12. What would the missing number in  $\_\_ + 6$  have to be for this expression to also equal 12? Since I know  $6 + 6$  is equal to 12, 6 must be the correct answer.” However a student who sees the equals sign as a command to do the indicated operation, might reason as follows: “It says to add  $7 + 5$ . So when I do the equals command to add them, I get 12 for the missing number.” Because students who see the equals sign as an operator would obtain a different answer than those who understand its meaning, this would be a good problem to assess whether students understand the mathematically correct meaning of the equals sign.

**Option C:**

Students who understand the equals sign as a command to carry out the operation could still get a correct answer for this problem. They could simply add the numbers on the right hand side of the equals sign to find the answer for the missing number. This is not a good choice to assess whether students understand the mathematically correct meaning of the equals sign.

**Option D:**

This option is similar to Option A and Option C above. Students who use the equals sign as a command to carry out the operation could still get a correct answer for this problem by adding the numbers in either of the expressions to find the answer for the missing number. This option is not a good choice to assess whether students understand the mathematically correct meaning of the equals sign.

Option B is the best answer, because students who see the equals sign as an operator would obtain a different answer from those who understand its meaning; this would be a good problem to assess whether students understand the mathematically correct meaning of the equals sign.

Summary of key knowledge, skills, and reasoning

This assessment draws on the following:

- Knowledge about the difference between an operational and an equivalence view of the equals sign.
- Knowledge that an operational view of equivalence can lead to viewing the equals sign as a command to do the math indicated in an expression.
- The ability to apply an operational view of equals to the problems in order to determine whether they produce different numerical results.
- Understanding that a good assessment problem should reveal common student mistakes to the teacher—a student whose reasoning is incorrect should answer incorrectly, so that the teacher knows a mistake has been made.

## Task Design Rationale – Lee: Fraction Comparison

Mr. Lee asked his students to compare  $\frac{7}{8}$  and  $\frac{6}{9}$ . All of his students correctly answered that  $\frac{7}{8}$  is greater than  $\frac{6}{9}$ , but they offered a variety of responses when asked to explain their reasoning. Of the following, which student responses provide mathematically valid explanations for why  $\frac{7}{8}$  is greater than  $\frac{6}{9}$ ? For each student response, indicate whether or not it provides a mathematically valid explanation.

	Provides a Mathematically Valid Explanation	Does Not Provide a Mathematically Valid Explanation
When you compare them, $\frac{7}{8}$ is greater than $\frac{6}{9}$ because 7 is greater than 6.	<input type="checkbox"/>	<input type="checkbox"/>
You can see that $\frac{7}{8}$ is greater than $\frac{6}{9}$ because ninths are smaller than eighths, which means that $\frac{6}{9}$ is less than $\frac{6}{8}$ which is less than $\frac{7}{8}$ .	<input type="checkbox"/>	<input type="checkbox"/>
You just need to look at how many pieces are missing. $\frac{7}{8}$ is greater than $\frac{6}{9}$ because $\frac{7}{8}$ is only missing one piece from the whole, but $\frac{6}{9}$ is missing three pieces from the whole.	<input type="checkbox"/>	<input type="checkbox"/>
I think $\frac{7}{8}$ is greater than $\frac{6}{9}$ because $\frac{7}{8}$ has more pieces than $\frac{6}{9}$ and those pieces are larger.	<input type="checkbox"/>	<input type="checkbox"/>
$\frac{7}{8}$ is greater than $\frac{6}{9}$ because $\frac{6}{9}$ is equal to $\frac{2}{3}$ , and because $\frac{1}{3}$ is greater than $\frac{1}{8}$ , $\frac{2}{3}$ is farther away from 1 than $\frac{7}{8}$ is.	<input type="checkbox"/>	<input type="checkbox"/>

### What is this assessment task asking?

In this task, Mr. Lee has asked his students to compare two fractions. Your task is to judge the mathematical validity of each of five separate student responses. You will read through samples of student work/reasoning and judge whether or not the particular student is demonstrating a valid reasoning process. In this task, validity means that the method is a correct way of obtaining the answer to this problem.

### What information is important?

It is important to pay attention to each student's reasoning independent of one another. It is important to notice that in some cases, all of the reasoning might not be explicit. You might use a certain amount of inference in your evaluation of the student work.

### What is the rationale for selecting an answer?

#### **Row A:**

You should notice that the student has based his or her answer on only a comparison of numerators. This is clearly not valid as a method except in the special case that the two fractions have equal denominators (and these do not), as it is the ratios that are being compared. A counterexample could easily be produced where a student might incorrectly identify the smaller fraction as the one with the smaller numerator ( $\frac{3}{5}$  and  $\frac{2}{3}$ , for example). In Row A, the student coincidentally arrived at the correct answer, but the reasoning is not mathematically valid.

**Row B:**

The student starts by stating that ninths are smaller than eighths. The student uses this reasoning to conclude that  $\frac{6}{9} < \frac{6}{8}$  (because it has an equal number of smaller pieces). It is also correct to conclude that  $\frac{6}{8} < \frac{7}{8}$  (because it has fewer pieces of the same size). Implicitly applying the transitive property for inequalities (if  $a < b$  and  $b < c$ , then  $a < c$ ), the student concludes that  $\frac{6}{9} < \frac{7}{8}$ . Although much of the reasoning is absent from this student's work, the work that is shown provides sufficient evidence of mathematically valid reasoning.

**Row C:**

The student has considered the number of “missing pieces” of each fraction. This can be a reasonable method to use, but the student has not accounted for the sizes of the missing pieces, only the quantity of them. A counterexample can easily be produced—for example,  $\frac{2}{3}$  is not greater than  $\frac{97}{100}$  even though  $\frac{97}{100}$  has more missing pieces (In this case, the three missing pieces in  $\frac{97}{100}$  are very small compared to the one missing piece of  $\frac{2}{3}$ ). For this reason, this student's work does not provide evidence of mathematically valid reasoning.

**Row D:**

It is true as stated that  $\frac{7}{8}$  has more pieces (greater numerator) than  $\frac{6}{9}$  and that those pieces are larger (smaller denominator). Combining a comparison of both numerators and denominators in this way is a mathematically valid way to reason about the comparison of fractions.

**Row E:**

The student begins by stating that  $\frac{6}{9}$  is equivalent to  $\frac{2}{3}$ . By using this equivalence, it can be seen that  $\frac{7}{8}$  and  $\frac{2}{3}$  are each a unit fraction ( $\frac{1}{8}$  and  $\frac{1}{3}$ , respectively) away from 1 or a whole. Since  $\frac{1}{3}$  is greater than  $\frac{1}{8}$ , then  $\frac{2}{3}$  must be farther from 1 than  $\frac{7}{8}$  is. (One way to think of this is that the student may have correctly applied the missing piece methodology suggested in Row C by taking only one piece from each fraction and comparing the relative sizes.) It is not clearly explained how the student knows that  $\frac{1}{3}$  is greater than  $\frac{1}{8}$ , but it is reasonable to assume, based on the student's demonstrated reasoning, that the student is able to compare fractions with like

numerators, or that the student may be familiar with these as benchmark fractions. This method is a mathematically valid explanation of why  $\frac{6}{9} < \frac{7}{8}$ .

Rows B, D, and E provide evidence of valid mathematical explanations, while Rows A and C do not.

Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Ability to evaluate student work for evidence of correct mathematical reasoning.
- Familiarity with common student errors with fractions such as comparing only numerators and/or denominators.
- Familiarity with common student errors with fractions such as comparing missing pieces.

## Task Design Rationale – Franco: Fraction Ordering

Ms. Franco was assessing students' work on comparing fractions. She assigned the following problem.

Put the following fractions in increasing order and explain your reasoning.  $\frac{4}{7}$ ,  $\frac{5}{8}$ ,  $\frac{2}{5}$

She noticed that Zachary got a correct answer with incorrect reasoning.

He explained that  $\frac{2}{5} < \frac{4}{7} < \frac{5}{8}$  because  $2 < 4 < 5$  and  $5 < 7 < 8$ .

To help Zachary understand that his reasoning is incorrect, Ms. Franco wants to give a similar problem using 3 different fractions. She wants to include fractions with 3 different numerators and 3 different denominators that, using Zachary's reasoning, would lead to ordering the fractions incorrectly, from greatest to least instead of least to greatest. List 3 such fractions in the boxes below in any order.

<input type="text"/>	<input type="text"/>	<input type="text"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>

### What is this assessment task asking?

This assessment task is asking you to find a set of fractions that serves as a counterexample for faulty student reasoning. This task requires you to think simultaneously about the relationships between the values of fractions and the values of their numerators and denominators and to be able to identify a set of fractions that satisfies multiple conditions. In particular, you are looking for three fractions that, when ordered using the student's method, will result in an incorrect ordering of the fractions.

The student in the assessment task has used the following incorrect reasoning:

Given three fractions,  $\frac{n_1}{d_1}$ ,  $\frac{n_2}{d_2}$ , and  $\frac{n_3}{d_3}$ , when the fractions are arranged in increasing order of the numerators and denominators, that is,  $n_1 < n_2 < n_3$  and  $d_1 < d_2 < d_3$ , then the fractions themselves will be in increasing order, that is,  $\frac{n_1}{d_1} < \frac{n_2}{d_2} < \frac{n_3}{d_3}$ .

The assessment task is asking you to find three fractions such that when the fractions are arranged in increasing order of the numerators and denominators, that is,

$n_1 < n_2 < n_3$  and  $d_1 < d_2 < d_3$ , then the fractions themselves will be in decreasing order,

$$\frac{n_1}{d_1} > \frac{n_2}{d_2} > \frac{n_3}{d_3}.$$

### What information is important?

There are a number of things you need to keep in mind when generating your three fractions:

- The three numerators must be different.
- The three denominators must be different.

- When the fractions are arranged by increasing numerators, then denominators must also be increasing. (Likewise, if the fractions are arranged by increasing denominators, then the numerators must also be increasing.)
- When the fractions are arranged by increasing numerators and denominators, the fractions themselves will be in decreasing order—not just any order other than increasing.
- The three fractions can be entered in the response boxes in any order.
- It is also important to notice here that although you are not asked for a general method for finding such a set of fractions, it is much easier to do if you have a systematic strategy for finding them rather than guessing and checking until you find a set that works.

What is the rationale for selecting answer?

A correct answer is one that meets the given conditions. When arranged by increasing numerators, the denominators must also be in increasing order and the fractions themselves must be in strictly decreasing order. The fractions need not be proper fractions and need not be in fully reduced form. Any set of fractions that meets these conditions forms a correct answer, but an important part of the task is not just whether one finds an answer but *how* one finds an answer.

There are many different reasoning paths that could be taken to generate or select appropriate examples, and there are an infinite number of examples that could be chosen with the desired relationships. A few of the methods that could be used to generate the examples are described below.

One method is to use common benchmark fractions in your reasoning process. This method should be accessible to students as well. You might start with  $\frac{3}{4}$ , which is equivalent to 0.75 but that has a relatively small numerator and denominator. Your next fraction must have a numerator greater than 3, a denominator greater than 4, but an overall value that is less than  $\frac{3}{4}$ . You might then choose  $\frac{5}{9}$  as your second fraction since it is a little greater than  $\frac{1}{2}$ . Finally, the third fraction could be  $\frac{6}{13}$ , as  $6 > 5$ ,  $13 > 9$ , and  $\frac{6}{13}$  is a little less than  $\frac{1}{2}$ . By using comparisons to common benchmark fractions, one can see that these three fractions would be ordered from greatest to least when the numerators and denominators are ordered from least to greatest.

A second method is to start with three fractions that are equal:  $\frac{1}{2} = \frac{2}{4} = \frac{3}{6}$ . You can then increase the denominators of the second and third fraction so that they lessen the overall value of the fractions:  $\frac{1}{2} > \frac{2}{5} > \frac{3}{10}$ .



A third method is to begin with decimals that one can easily identify as meeting the conditions:  $0.1 > 0.02 > 0.003$  (in fractional form:  $\frac{1}{10} > \frac{2}{100} > \frac{3}{1000}$ ) even though  $1 < 2 < 3$  and  $10 < 100 < 1000$ .

A fourth method uses the fact that the assessment task does not state that the three fractions must each have a value less than 1. So consider any fraction greater than 1,  $\frac{n}{d}$  where  $n > d$ , and the two fractions formed by adding 1 and 2 to both the numerator and denominator,  $\frac{n+1}{d+1}$  and  $\frac{n+2}{d+2}$ . You can show that  $\frac{n}{d} > \frac{n+1}{d+1} > \frac{n+2}{d+2}$ , provided that none of the denominators equal 0, thus generating a set of fractions that meet the desired criteria. For example, start with  $\frac{7}{3}$ . Generate the next two fractions,  $\frac{8}{4}$  and  $\frac{9}{5}$ . It is true that the numerators are increasing,  $7 < 8 < 9$ , and the denominators are increasing,  $3 < 4 < 5$ , but the fractions are decreasing,  $\frac{7}{3} > \frac{8}{4} > \frac{9}{5}$ .

You can see from this handful of examples that there are quite a few systematic approaches to this assessment task that will result in correct solutions.

#### Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Knowledge of ordering fractions.
- Ability to construct a set of fractions that meet a given set of conditions.
- Understanding of the concept of a counterexample.

## Task Design Rationale – Hupman: Simple Exponential Expressions

Ms. Hupman is teaching an introductory lesson on exponents. She wants to give her students a quick problem at the end of class to check their proficiency in evaluating simple exponential expressions. Of the following expressions, which would be least useful in assessing student proficiency in evaluating simple exponential expressions?

- $3^3$
- $2^3$
- $2^2$
- All of these are equally useful in assessing student proficiency in evaluating simple exponential expressions.

### What is this assessment task asking?

What this assessment task asks you to do is to choose the least useful of a set of problems for assessment purposes, but without specifying what the criteria for usefulness should be. In order to answer, you need to understand that a useful problem for assessing students is one that reveals to the teacher whether or not the students have understood. Choosing which of these is least useful requires thinking through what a student misunderstanding is likely to look like (what mistake a student might make) and then determining which of these problems would reveal that mistake to the teacher (making them useful for assessment) and which would obscure the misunderstanding.

### What information is important?

While there are many mistakes that students might make in evaluating exponential expressions, a common one is multiplying the base by the exponent. Other less common errors include reversing the base and the exponent or choosing other operations (reading it as addition, for example). You may be familiar with these errors, or you may be able to think about the given exponential expressions and imagine what the possibilities are for evaluating incorrectly. A useful problem for assessment would be one that would alert the teacher if a student uses one of these incorrect methods. A less useful assessment problem would be one that the student could coincidentally answer correctly using one of these erroneous methods, concealing the problem from the teacher.

### What is the rationale for selecting an answer?

#### **Option A:**

If  $3^3$  is evaluated incorrectly by multiplying base times exponent, the result is  $3 \times 3 = 9$ , which is a different answer from the correct answer,  $3 \times 3 \times 3 = 27$ , so the teacher would know that an error has been made.

If evaluated incorrectly by reversing the exponent and the base, the result would be 27, which is the same as the correct answer, so the teacher might not know that an error has been made. If evaluated incorrectly by adding the base and exponent, the result would be 6, which is different from the correct answer, so the teacher would know that an error has been made.

The most common mistake is revealed by this problem, but one other possible mistake is not, so this problem is only somewhat useful in revealing student errors.

**Option B:**

If  $2^3$  is evaluated incorrectly by multiplying the base times the exponent, the result is  $2 \times 3 = 6$ , which is different from the correct answer,  $2 \times 2 \times 2 = 8$ , so the teacher would know that an error has been made.

If evaluated incorrectly by reversing the exponent and base, the result would be 9, which is also different from the correct answer and would reveal the error. If evaluated incorrectly by adding the base and exponent, the result would be 5, which is again different from the correct answer.

This is a useful assessment problem that reveals several common errors.

**Option C:**

If  $2^2$  is evaluated incorrectly by multiplying base times exponent, the result is  $2 \times 2 = 4$ , which is the same as the correct answer,  $2 \times 2 = 4$ , and would conceal the mistake from the teacher.

If evaluated incorrectly by reversing the exponent and base, the answer would be 4, which is the same as the correct answer, and likewise would conceal the mistake. If evaluated incorrectly by adding the base and exponent, the answer would still be 4 and the teacher would not know an error had been made. This assessment problem hides several common errors from the teacher by allowing the student to arrive at a correct answer using incorrect reasoning, and it is less useful than the other options.

**Option D:**

Option D is not the best answer because there are clear differences in usefulness among the options.

Option C, because it would conceal several common errors from the teacher, is the least useful for assessing student proficiency in evaluating simple exponential expressions.

Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Awareness that, to be useful, an assessment problem should reveal common student mistakes to the teacher—a student whose reasoning is incorrect should answer incorrectly so that the teacher knows a mistake has been made.
- Awareness that students commonly make certain errors in calculating exponential expressions, including multiplying the base by the exponent.

## Task Design Rationale – Richmond: Proportional Reasoning

In a unit on proportional reasoning, Ms. Richmond's class was discussing the following problem.

If 4 cups of cocoa and 2 cups of sugar yield 16 brownies, how many cups of cocoa and how many cups of sugar are needed to make 24 brownies?

Ms. Richmond's students used different strategies to solve the problem. For each strategy, indicate whether or not it provides evidence of mathematically valid student thinking.

	Provides Evidence of Mathematically Valid Student Thinking	Does Not Provide Evidence of Mathematically Valid Student Thinking
48 brownies need 12 cups of cocoa and 6 cups of sugar. To make 24 brownies, I need 6 cups of cocoa and 3 cups of sugar.		
4 and 2 both go into 16. 4 plus 2 is 6, half of 6 is 3, and 6 and 3 both go into 24, so you need 6 cups of cocoa and 3 cups of sugar to make 24 brownies.		
1 brownie needs $\frac{1}{4}$ cup of cocoa and $\frac{1}{8}$ cup of sugar. To make 24 brownies, I need to multiply by 24 for cocoa and sugar. Thus, I need 6 cups of cocoa and 3 cups of sugar.		
6 cups of cocoa and sugar makes 16 brownies, so 24 brownies need 9 cups of cocoa and sugar. Since the ratio of cocoa to sugar is 2:1, I need 6 cups of cocoa and 3 cups of sugar.		
Since 1 cup of sugar is needed to make 8 brownies, I need 3 cups of sugar to make 24 brownies. The amount of cocoa is two times the amount of sugar in the recipe, so I need 6 cups of cocoa.		

### What is this assessment task asking?

In this assessment task, you'll need to evaluate each student explanation separately and decide whether there is enough evidence to conclude that the method shown is mathematically valid. All of the students have arrived at correct answers, so the point is to evaluate the methods, not the solutions. The main work of this assessment task can be thought of, then, as having two steps. First, figure out what general method is implied from the student explanation. Second, decide whether that method is valid. Because what is asked for is "evidence," an explanation does not have to be completely clear or concise to qualify; there just has to be a plausible, valid way of understanding what the student was probably thinking.

### What information is important?

It is possible to answer this assessment task by going straight to the options and considering each in turn, but it helps to notice a few things about what is going on in the original problem first. The problem involves three quantities (amount of cocoa, amount of sugar, and number of brownies) that should remain in fixed proportionality relative to one another no matter how the batch size is altered. Therefore, "mathematically valid" student thinking should involve manipulating these quantities in a way that maintains their relative proportionality. This can be done by scaling all three quantities at once (for example, an efficient solution is to multiply all three quantities by 1.5, although this is not a likely method for students to use) or by working with them in pairs, as long as the third quantity is accounted for afterward.

## What is the rationale for selecting an answer?

### **Row A:**

The student whose explanation is in Row A says that 48 brownies require 12 cups of cocoa and six cups of sugar, which is a true statement. The second statement, that you need six cups of cocoa and three of sugar to make 24 brownies, is also true. However, the explanation does not make clear how the student arrived at the first statement or how the student moved from the first statement to the second. Figuring out what the student was thinking requires a little imagination. It seems reasonable that the student tripled the given recipe to get a recipe for 48 brownies and then halved that one to get a recipe for 24 brownies. In doing so, the student provides a valid solution process, since scaling (by a factor of 3 and then by a factor of  $\frac{1}{2}$ ) preserves proportionality. Thus, there is evidence here of mathematically valid student thinking.

### **Row B:**

The explanation in Row B is a series of true statements with little connective tissue. It is true that 4 and 2 both go into 16, that 4 plus 2 is in fact 6, and that half of 6 is 3, but the student is not explicit about why this matters. Trying to imagine what the student might have been thinking, it seems reasonable to assume that the 4 and 2 are the cups of cocoa and sugar from the original recipe and that the 6 probably means 6 cups of cocoa/sugar mix, and in this case, 3 would represent half the mix. The numbers 6 and 3 do indeed each go into 24, but it is not clear why their being factors of 24 would be a reason to expect that they would then represent the desired quantities of cocoa and sugar. In fact, if this line of reasoning does indeed represent this student's thinking, the same amount of cocoa and sugar would have been required for batches of 6, 12, 18, 24...brownies (because in each case, 6 and 3 divide into the number of brownies). There does not seem to be any evidence here of mathematically valid thinking.

### **Row C:**

The student whose explanation is in Row C states that one brownie needs  $\frac{1}{4}$  cup cocoa and  $\frac{1}{8}$  cup sugar, and it is reasonable to imagine the student reached this conclusion by dividing each quantity by 16 as a means of finding the per-unit (brownie) ingredient amounts. The student has been explicit that from there she multiplied the unit ingredients by the desired number of brownies to arrive at the solution. Scaling by a factor of  $\frac{1}{16}$  and then by a factor of 24 preserves proportionality, so the method is valid, and this row has evidence of mathematically valid student thinking.

### **Row D:**

The statement that six cups of cocoa and sugar makes 16 brownies is unclear—it may mean six cups each of cocoa and sugar (which would be incorrect) or it may mean six cups combined from the four cups of cocoa and two cups of sugar (which would be correct). Assuming the second

interpretation, it is also true that 24 brownies need nine cups of ingredients, although the student did not specify how this was decided. The student states explicitly how the final cocoa and sugar amounts are calculated by breaking nine cups into two parts with a ratio of 2:1, and this reading supports our assumption that the first statement was referring to six combined cups of cocoa and sugar. This student seems to think of this problem as (mix):(brownies) and then as (cocoa):(sugar) within the mix. Scaling from six to nine cups of mix preserves proportionality, and the last step explicitly preserves the proportionality of cocoa to sugar, so this method is also valid, even though several steps of the process are not clearly explained. While there is less explanation here than in some of the other rows, there is certainly evidence of mathematically valid student thinking.

### **Row E:**

This student states that one cup of sugar makes 8 brownies. This is correct, and a reasonable way the student might have reached this conclusion is to divide the given quantities by two. Similarly, it is correct that three cups of sugar makes 24 brownies and reasonable that the student would reach this conclusion by multiplying both quantities by three. It is also correct that once having found the amount of sugar, the student can calculate the amount of cocoa by maintaining the constant ratio 2:1, as the student explains. Like the student in Row D, this student has chosen to consider things pairwise by looking at sugar in isolation first, and then going back to calculate the amount of cocoa. This student has scaled by a factor of  $\frac{1}{2}$  and then by a factor of 3 correctly and has maintained the relative ratio of cocoa to sugar correctly, so the method is valid. There is evidence here of mathematically valid student thinking.

Rows A, C, D, and E all demonstrate evidence of mathematically valid student thinking, but Row B does not.

### Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Understanding of proportionality, including the knowledge that three quantities can maintain a proportional relationship simultaneously and that scaling quantities (to make them larger or smaller) preserves proportionality.
- Ability to recognize various general approaches to the problem or to fill in the pieces in the underlying mathematics to create a coherent method from the incomplete explanations provided.
- Ability to coordinate the two processes of creating coherent reasoning and evaluating on the basis of mathematical validity.

## Task Design Rationale – Santiago: Order of Operations

To assess her students' prior knowledge about evaluating arithmetic expressions, Ms. Santiago assigned a worksheet of problems. She noticed that Alexis answered the first two incorrectly and the next two correctly.

$$\begin{array}{l} 1) 7 \times 2 - 6 + 3 = 5 \\ 2) 9 - 5 + (16 \div 8) = 2 \\ 3) 9 + 24 \div 3 - 1 = 16 \\ 4) 17 - (3 + 7 \times 2) = 0 \end{array}$$

Which of the remaining problems is Alexis likely to answer incorrectly?

- $8 + 7 - 12 \div 3$
- $13 - 3 \times 2 + 5$
- $(27 \div 3 - 4) + 8$
- $(16 - 12) \times 5 + 10$

### What is this assessment task asking?

Although this assessment task asks you to identify which problems the student is likely to answer incorrectly, the primary challenge is figuring out what Alexis is doing based on the work samples that are given. This means first figuring Alexis' source of confusion as demonstrated by the combination of correct and incorrect work. The next step is determining how her confusion might lead to answering incorrectly for each of the four problem choices. Answering this assessment task is aided by the knowledge that students who have learned to solve problems by drawing on mnemonics used to help them remember the order of operations are likely to make the same errors that Alexis demonstrates in her work.

### What information is important?

To answer this assessment task, you first need to analyze the four examples of Alexis' work. You need to understand what she did to get the first two problems wrong and then test your hunch about her confusion to see whether it is consistent with answering the other problems correctly.

In the first example, Alexis should have done the following:

$$\begin{array}{l} 7 \times 2 - 6 + 3 \quad \text{multiply 7 by 2} \\ = 14 - 6 + 3 \quad \text{subtract 6 from 14} \\ = 8 + 3 \quad \text{add 8 and 3} \\ = 11 \end{array}$$

However, Alexis gave 5 as an answer instead of 11. What might she have done that would explain this?

One way of combining the numbers incorrectly that leads to a result of 5 is shown below:



$$\begin{aligned}
& 7 \times 2 - 6 + 3 \quad \text{multiply 7 by 2} \\
& = 14 - 6 + 3 \quad \text{add 6 and 3} \\
& = 14 - 9 \quad \text{subtract 9 from 14} \\
& = 5
\end{aligned}$$

There may be other ways of changing the expression that lead to a result of 5. For example, she might have misread the plus sign between 6 and 3 as a minus sign. However, this misreading, while possible, is less likely than confusing the order of operations as illustrated above.

At this point, we have an idea of what happened on the first problem, but one example is not enough. We still can't choose from the range of reasons that could explain why she made the error. Is she "chunking" the expression—doing the left hand part and the right hand part separately and then combining with the middle operation? Is she adding before subtracting? Is she inserting parentheses in ways that we don't yet understand?

In the second example, Alexis should have done the following:

$$\begin{aligned}
& 9 - 5 + (16 \div 8) \quad \text{divide 16 by 8} \\
& = 9 - 5 + 2 \quad \text{subtract 5 from 9} \\
& = 4 + 2 \quad \text{add 4 and 2} \\
& = 6
\end{aligned}$$

However, she arrived at 2 as a result instead of 6. What might she have done that would explain this?

$$\begin{aligned}
& 9 - 5 + (16 \div 8) \quad \text{divide 16 by 8} \\
& = 9 - 5 + 2 \quad \text{add 5 and 2} \\
& = 9 - 7 \quad \text{subtract 7 from 9} \\
& = 2
\end{aligned}$$

There is a pattern emerging: Alexis seems in both cases to have added and then subtracted where she ought to have subtracted and then added. But again, we still do not know why she made the error. It seems less likely now that she might be chunking, or dividing the problem into left and right sides, since doing so would have yielded a correct answer on Example 2. She could be adding before subtracting. While there is evidence here that she uses parentheses correctly, we still don't know exactly why she divided first. Was it because of the parentheses or because the operation was division?

In the third example, Alexis solved correctly:

$$\begin{aligned}
& 9 + 24 \div 3 - 1 && \text{divide 24 by 3} \\
& = 9 + 8 - 1 && \text{add 9 and 8} \\
& = 17 - 1 && \text{subtract 1 from 17} \\
& = 16.
\end{aligned}$$

She also could have subtracted 1 from 8 and then added 9 and 7, but in this case it happens to make no difference. It is also less likely, given her inclination to add before subtracting up to this point.

What do we know from the third problem? The emerging pattern of addition before subtraction holds (here it does not happen to be incorrect to add before subtracting). We also know that she divided before adding or subtracting, even in the absence of parentheses.

In the fourth example, Alexis solved correctly:

$$\begin{aligned}
& 17 - (3 + 7 \times 2) && \text{multiply 7 by 2} \\
& = 17 - (3 + 14) && \text{add 3 and 14} \\
& = 17 - 17 && \text{subtract 17 from 17} \\
& = 0.
\end{aligned}$$

You can also use familiarity with common student misconceptions to help think about and answer this assessment task. Order of operations is often taught using the mnemonic PEMDAS, which is often referred to by a label such as “Please Excuse My Dear Aunt Sally.” This mnemonic is meant to indicate that one should complete, in order, parentheses, exponents, multiplication/division, and addition/subtraction. The use of PEMDAS is handy but can lead to a number of misconceptions. A common misconception is that the operations go strictly in order as listed—that is, that multiplication always comes before division and that addition always comes before subtraction. Since we know that Alexis is adding before subtracting, it is likely that this is the underlying cause of the errors. It was not necessary to know this information to solve the problem, but familiarity with this common error would make it easier to figure out what Alexis is doing.

What can you conclude about Alexis’s thinking?

- Alexis correctly does multiplication and division before either addition or subtraction.
- We have some evidence that Alexis correctly does work in the parentheses first, although this is not conclusive.
- Alexis incorrectly does addition before subtraction, except in cases where the subtraction is in parentheses.
- Alexis’s work suggests she may be relying on an incorrect understanding of a mnemonic such as PEMDAS to solve the problems.
- There is no apparent pattern of incorrectly chunking the work.

### What is the rationale for selecting an answer?

At this point, we know the likely cause of the error (addition before subtraction) and must decide which of the options would be answered incorrectly if this pattern were continued for the next four problems.

#### **Option A:**

In Option A, addition before subtraction would give a correct answer, so Alexis is not likely to answer incorrectly.

#### **Option B:**

In Option B, if Alexis continues the same pattern of doing addition before subtraction, she will answer incorrectly. Addition before subtraction is incorrect for this problem, and there are no parentheses to guide her as there are in Option C. She will likely answer incorrectly.

#### **Option C:**

In Option C, if we assume that Alexis does the work in parentheses first, she would answer correctly.

#### **Option D:**

In Option D, once again, if Alexis correctly uses parentheses, she would get the right answer.

Option B is the one that Alexis is most likely to answer incorrectly, and it is the best option.

### Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Knowledge of how order of operations is used to evaluate numerical expressions.
- Familiarity with common student errors with order of operations, including confusing the order of operations and incorrect use of strategies such as chunking.
- Awareness that tools, methods, and other aids that are commonly taught to help students solve problems can also support student misconceptions.
- Ability to identify problems that are likely to reveal a student error or source of confusion.
- Ability to analyze student work to identify the steps that were used to arrive at correct and incorrect solutions.

## Task Design Rationale – Sabatine: Area and Perimeter

At the start of a lesson on finding the side length of a square given its area, Ms. Sabatine gave her students a problem to assess their prior knowledge. Several students incorrectly answered that the side length of a square with area 36 square units is 9 units. At the end of the lesson, Ms. Sabatine wanted to give a similar problem to assess what her students had learned. Of the following areas to use in this problem, which would be least useful for assessing student learning in this situation?

- 9 square units
- 16 square units
- 64 square units
- 100 square units

### What is this assessment task asking?

What this assessment task asks you to do is to choose the least useful of a set of problems for assessment purposes, where the meaning of *useful* is tied to the classroom context. A particular student error is described, and the teacher wants to assess the students at the end of the lesson to see whether that particular misconception has been addressed. In order to answer, you need to understand that a useful assessment problem is one that reveals to the teacher whether or not the student continues to hold the described misconception. Which is least useful can be determined by considering, for each option, whether it gives the same numerical answer worked correctly as it does worked incorrectly. Problems that allow a correct answer via incorrect reasoning hide that incorrect reasoning from the teacher, so they are not as useful for assessment.

### What information is important?

To answer this assessment task, you first need to note the error that the students are making. The students here answered that the side length of a square with an area of 36 square units is 9 units. (The correct answer is 6.) Why the students did this is not clear, but if you are familiar with common student errors, you may realize that the students are probably confusing area and perimeter. (If the perimeter of a square were 36 units, each side would have length of 9 units.) Another possibility is that they are overlooking that the figure is a square. (If a rectangle had an area of 36 square units, one possible shape would be 4 by 9.) You do not need to recognize the reason for the error to answer the assessment task, but doing so may make it easier to imagine what the students would do with the problems described in the options.

### What is the rationale for selecting an answer?

#### **Option A:**

For Option A, the side length, if calculated correctly, would be 3 units, and if calculated incorrectly by the given method (dividing by 4), would be 2.25 units. However, the fact that the mistake is computationally more difficult than the correct method might lead

students to answer correctly for the wrong reason—because the square root of 9 is easier to calculate than 9 divided by 4. The answers do not match, so the teacher would know an error had been made, but some students who still do not understand might be led away from the error. This is a somewhat useful problem for assessing student learning.

**Option B:**

For Option B, the side length, if calculated correctly, would be 4 units, and if calculated incorrectly by the given method (dividing by 4), would be 4 units. The answers are the same, so the teacher would have no way of knowing the student had made an error. This is not a useful problem for assessing student learning.

**Option C:**

For Option C, the side length, if calculated correctly, would be 8 units, and if calculated incorrectly by the given method (dividing by 4), would be 16 units. The answers do not match, so the teacher would know an error had been made, making this a useful problem for assessing student learning.

**Option D:**

For Option D, the side length, if calculated correctly, would be 10 units, and if calculated incorrectly by the given method (dividing by 4), would be 25 units. The answers do not match, so the teacher would know an error had been made, making this a useful problem for assessing student learning.

Option B is the only one in which a student could coincidentally arrive at a correct answer using the incorrect reasoning described in the assessment task, and this makes it the least useful for assessing student learning.

Summary of key knowledge, skills, and reasoning

- To be useful, an assessment problem should reveal student mistakes to the teacher—a student whose reasoning is incorrect should answer incorrectly so that the teacher knows a mistake has been made.
- Students commonly make certain errors in calculating area and perimeter, and a common such mistake is using the computations associated with one when the other is called for.

## Appendix B

### Task Design Rationales for ELA Items

#### Task Design Rationale – Kumar: Referents Causing Reading Comprehension Difficulties

While students are reading independently, Ms. Kumar begins talking with Doug, a struggling fourth-grade reader, about the following two sentences he just read from a book about a girl and her dog.

I have a dog named Glover. That mischievous mutt is always getting into trouble.

Doug comments that his dog is bad too, and that he is excited to read more about the two dogs in the story. Ms. Kumar is trying to figure out why Doug is misreading the text.

Which of the following is the best explanation for Doug's confusion?

- Doug has not comprehended the word "mischievous."
- Doug has made a connection to his dog, which has interfered with his comprehension.
- Doug does not understand that "Glover" and "mischievous mutt" refer to the same dog.
- Doug is confused by the complex sentence structures used in the text.

#### What is this assessment task asking?

This assessment task is asking you to analyze a short two-sentence passage and make a judgment about what features of the passage most likely led to a student's confusion. The analysis requires that you pay attention to and use a range of information about the student, Doug, including the confusion indicated by his comments, as well as information about typical patterns of confusion for students like Doug. Responding to this assessment task is aided by the knowledge that struggling readers often have trouble keeping track of referents that appear across a text passage, including text passages as short as two sentences.

#### What information is important?

To respond to this assessment task, you need to first make sense of what Doug has demonstrated in his comment about the text passage. Doug has clearly read the two sentences and understood key information. He recognizes that this is a story about a dog getting into trouble. He draws a connection to his own dog and comments that his dog is also bad. The passage has caught Doug's interest, and he is clearly engaged and wants to read more. However, Doug is also confused. He thinks that this is a story about "two dogs," not a single "mischievous mutt" as stated in the text.

Deciding on what could be causing Doug's confusion requires drawing on information about Doug and his reading. Doug is a struggling reader. The text passage is somewhere between late first- and second-grade level. This indicates that Doug is likely to have difficulties that are not demonstrated by fluent fourth- or even third-grade readers. Expected difficulties could include problems such as not reading a text for meaning, lack of familiarity with text content, or

confusion over key concepts or vocabulary. However, as noted above, Doug’s comments about the text suggest that he is reading the text for meaning and is making a strong connection to his own experience. Doug has a good grasp of the basic concepts in the passage, including recognition that this is about a dog that gets into trouble.

The text itself has a simple structure with few clauses or unusual use of verb tense, both features of text that could lead to comprehension difficulties. The word *mischievous* is likely to be unfamiliar to a struggling reader. However, Doug’s comment about his dog suggests that a misunderstanding of the word is not a source of his difficulty. Teachers who are experienced with struggling readers are likely to know from prior experience or research that confusion can arise around the use of referents in text. In this case, *dog*, *Glover*, and *mutt* all refer to the same dog.

What can you conclude about Doug’s reading?

- Doug understands that this is a story about a bad dog.
- Doug has shown that he has comprehended deeply enough to make a personal connection.
- Doug is interested in the story and motivated to read more.
- Doug is confused and thinks that this is a story about two dogs.

What is the rationale for selecting an answer?

In order to consider which of the four options is the best explanation for Doug’s misreading of the text, you must consider the four options in relation to each other.

**Option A:**

For a struggling fourth-grade reader, a single difficult vocabulary word such as *mischievous* can often interfere with comprehension. However, misunderstanding the word *mischievous* would not cause Doug’s particular misreading, since nothing about the word would lead a reader to think that there are two dogs in the story. Also Doug’s comment that “his dog is bad too” suggests that he may even know the meaning of *mischievous*. On the other hand, Doug may also have relied on the phrase “always getting into trouble” to help him make the connection to his bad dog. Therefore, Option A may not be the best answer.

**Option B:**

Option B describes something that Doug clearly did in his reading. Doug’s comment about his bad dog does show that he made a personal connection to the text, but because he mentions “the two dogs in the story,” his misreading does not seem to be related to his own dog. Therefore, Option B is not the best answer.

**Option C:**

Option C provides a reasonable explanation why Doug is confused. As noted above, “Glover” and “that mischievous mutt” are two references to the same dog. It is common for a struggling fourth grade reader to be confused by the use of more than one term to refer to the same person, place, or thing. Confusion about the relationship between pronouns (including the relative pronoun, *that*) and their antecedents is especially common. Doug could easily have interpreted these different ways of referring to the same dog as references to two separate dogs. Therefore, Option may be the best answer.

**Option D:**

Option D is not the best choice because, although syntax is often confusing for students learning to read, the sentences in the text are both simple sentences. Even for a struggling fourth-grade reader, they are not complex.

Option C is the best answer, since a struggling fourth-grade reader would often struggle to discern the relationships between pronouns and their antecedents.

**Summary of key knowledge, skills, and reasoning**

This assessment task draws on the following:

- Knowledge that subjects of sentences or passages can be referred to in a variety of ways.
- Knowledge of sentence complexity.
- Knowledge that it is not necessary to understand every vocabulary word to comprehend the meaning of a sentence (e.g., it is possible to understand this is a “bad dog” without knowing the meaning of the word *mischievous*).
- Ability to use knowledge of syntax and vocabulary to anticipate difficulties that a student could have reading a sentence.
- Knowledge that a reader will often interpret meaning in a text through the lens of his or her own personal experience, and this can help or hinder comprehension in a variety of ways.
- Knowledge that it is common for struggling readers in this age group to be confused by the use of more than one term to refer to a single subject



## Task Design Rationale – Haddad: Choosing Discussion Questions That Focus on Character Development

*Questions 7-9 are based on a lesson using Jerry Spinelli's novel "Maniac Magee."*

Mr. Haddad is using Jerry Spinelli's novel *Maniac Magee* to teach his fifth-grade class about how authors develop their characters.

Which of the following questions would be the best choice to help his students focus on the essential features of character development?

- What kinds of symbolism does the author use to develop Maniac Magee's character?
- Can you describe where Maniac Magee lives?
- How does Maniac Magee respond to trouble?
- What do you think might have happened to Maniac Magee before this story began?

### What is this assessment task asking?

This assessment task asks you to choose the best question for focusing a class discussion on character development in a novel. You don't need to have read *Maniac Magee* to respond to the assessment task correctly. However, you do need to be familiar with the essential features of character development so that you can identify which questions deal with these features. You may be able to identify the correct option through this knowledge alone, but in order to confirm your response, you should also determine whether the question you've chosen would *most directly* focus fifth-grade students' attention on an essential feature of character development.

### What information is important?

It is important to notice that the class is learning how authors develop their characters generally; this question is not focused on Spinelli's development of *Maniac Magee*. In order to discern the best answer, it is helpful to consider the universally identifiable traits that authors use to develop characters.

It is also helpful to notice that the assessment task asks you to choose a question that would help students to focus on the "essential features of character development." This means that the question has to be able to trigger discussion about the character's actions, speech, or appearance, or about perceptions of the character by others.

### What is the rationale for selecting an answer?

In order to consider which of the four options would be the best question for generating discussion about the author's character development in *Maniac Magee*, you must consider the four options in relation to each other.

**Option A:**

Symbolism is not an *essential* feature of character development. An author may use symbolism to add another layer of meaning to a text, but an effective author will strive to make sure that characters are developed in such a way that readers can understand who a character is, even if they do not catch the symbolism included in the text. Furthermore, symbolism is not usually covered in much detail in fifth grade, probably because most fifth graders aren't developmentally ready for the kind of abstract thinking it requires. Finally, if you have read *Maniac Magee*, you might also conclude that it doesn't contain a great deal of symbolism. Therefore, Option A is not the best answer.

**Option B:**

This question asks students to describe Maniac Magee's home and/or community rather than him as an individual. While some aspects of Maniac Magee's home and community may reveal certain aspects of his character, having students describe where he lives does not lead them directly to making inferences about what kind of person Maniac Magee is.

**Option C:**

Option C could easily lead students into a discussion of how the author develops Maniac Magee's character. By identifying how Maniac Magee responds to trouble in the text, students would be able to draw conclusions about what kind of person he is, incorporating evidence from the text about his actions and possibly his words and thoughts also. This answer choice is more directly focused on essential features of character development than Options A and B and is therefore a possible choice.

**Option D:**

Option D asks students to offer hypothetical scenarios about Maniac Magee's past, which, if done correctly, might be based upon character development information contained in the text, but would not demonstrate the work an *author* does to develop a character. This question may be an effective measure for assessing students' understanding of Maniac Magee's character but not for determining their understanding of how an *author* develops a character.

Option C is the best option because it has the potential to lead to a discussion of Maniac Magee's development as a character.

**Summary of key knowledge, skills, and reasoning**

This assessment task draws on the following:

- Knowledge that the essential features of character development include a character's actions, speech, and appearance, and perceptions of the character by others.

- Knowledge that symbolism may be used to supplement an author's development of a character, but it isn't generally an essential feature of character development, especially in novels for upper elementary readers.
- Knowledge that fifth graders may not be developmentally ready to deal with abstract concepts such as symbolism in their study of literature.
- Knowledge that an effective question to focus students' attention on an author's characterization techniques must be one that can be answered only with information from the text.

Knowledge that an effective question to focus students' attention on an author's characterization techniques should provide the most direct route to discussing the character, rather than approaching it indirectly.

## Task Design Rationale – Haddad: Supporting Student Completing Character Map Graphic Organizer

Questions 7-9 are based on a lesson using Jerry Spinelli's novel "Maniac Magee."

Mr. Haddad asks his students to create a character map—a graphic organizer in which students record words and short phrases that capture the most important aspects of a character's actions and memorable sayings, as well as how other people react to the character. Rachel has chosen John McNab, a minor character who appears infrequently in *Maniac Magee*. Mr. Haddad notices that Rachel has been copying every word in the novel about John McNab into her notebook.

Which of the following is the best teaching approach to help Rachel record useful information about her character's development?

- Interview her about what John McNab says and does at important points in the story.
- Suggest that she use a character map of Maniac Magee as a model for her own work.
- Show her how to discriminate between important and unimportant information in text.
- Help her create a list of events that involve John McNab, and model how to write about the first event.

### What is this assessment task asking?

This assessment task asks you to identify which teaching approach would best help Rachel to record useful information about a character's development in a novel. To respond to it correctly, you first need to pinpoint the reason why Rachel is having difficulty with the assignment. Once you have determined the cause of Rachel's difficulty, you can then evaluate the options to determine which option best addresses this cause. You don't need to have read *Maniac Magee* in order to respond to this task correctly.

### What information is important?

To respond to this assessment task, it is important to notice first what the assignment is. Rachel has been asked to fill out a character map graphic organizer in which she captures the most important aspects of a character's development by recording words and short phrases from the novel that describe his actions, his memorable sayings, and perception by other characters.

It is also significant that Rachel has chosen a minor character, John McNab; this limits the amount of information she can choose from when completing her map and may contribute to the trouble she is having with the assignment.

Rachel's approach to the assignment is probably the most critical element of the assessment task to notice. Instead of focusing just on important details, Rachel is copying down every word in the novel related to her chosen character, John McNab. This piece of context suggests that she is having trouble determining which details about John McNab are most important. She may also have trouble selecting only short words or phrases to capture these details. The first difficulty may be related to the fact that she has chosen a minor character. If there is little information about him in the first place, she may feel that she needs to capture it all.

### What is the rationale for selecting an answer?

At this point, we know that Rachel is probably filling out her graphic organizer incorrectly because she does not know how to distinguish which information about her character is important and which is not. Going forward, we need to select which of the options will most help Rachel record useful information. In order to consider which of the four options would best support Rachel in completing the graphic organizer successfully, you must consider the four options in relation to each other.

#### **Option A:**

This option would help Rachel record information from important parts of the story and would guide Rachel to record actions and memorable sayings, but it misses how other characters react to John McNab. Also, it is possible that there is relevant information in other parts of *Maniac Magee*; the important points of the story and the important points in John McNab's development may not necessarily be the same. For these reasons, Option A is not the best answer.

#### **Option B:**

Providing models for students is a common and effective way to scaffold their understanding of a task. However, if Rachel is struggling with identifying relevant details, the already completed map will not necessarily help her learn this skill, because she doesn't have to go through the process of choosing the details; they have already been chosen for her. For this reason, Option B is not the best answer.

#### **Option C:**

Discriminating between important and unimportant information is the skill that Rachel seems to be struggling with. If Mr. Haddad shows Rachel how to distinguish between important and unimportant information in the text, Rachel will likely be able to capture only the most important details relating to John McNab and will thus be able to fill out her character map accurately. Although this option does not specify how Mr. Haddad will accomplish this task, it forms a template for addressing Rachel's misunderstanding in a way that will help her resolve it. For these reasons, Option C may be the best answer.

#### **Option D:**

The first phase of this activity, creating a list of events in *Maniac Magee* that involve John McNab, does not help Rachel gather only useful information about this character. In fact, it may reinforce her habit of not distinguishing between important and unimportant information. If we assume that the second phase, modeling "How to write about the first event," means showing Rachel how to write about the event using the short words and phrases appropriate to a character map, then it does address part of the problem that Rachel is having; it demonstrates how to briefly capture the essence of an event rather than copying down every word about it. However,

if she still doesn't know how to determine which events involving John McNab reveal something important about his character, then she has missed the core goal of the assignment; in other words, if she's writing in the map about insignificant events, then it doesn't matter so much whether her comments about the events are brief. For these reasons, Option D is not the best answer.

Option C is the best answer because, unlike the other options, it addresses the main problem that Rachel is having and will help her correctly complete the character map assignment.

### Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Knowledge that creating a character map requires being able to distinguish between important and unimportant information in a narrative text and being able to select short words or phrases from the text that capture the essence of an idea.
- Knowledge that every appearance of a character in a novel doesn't necessarily provide essential information about his or her development.
- Knowledge that important events in a story are not always important points in a specific character's development.
- Knowledge that when students are asked to select important information from a text but instead copy it verbatim, this likely means that they don't understand how to distinguish between important and unimportant information.
- Knowledge that, when using modeling to support a struggling student, it is important to model the specific aspect of a task that is causing the student's trouble.

## Task Design Rationale – Goldberg: Generating Questions to Support Discussion of Passage Theme

*Questions 4-5 are based on the following teaching scenario.*

Mr. Goldberg is teaching the science fiction novel *The Giver*, by Lois Lowry, to his sixth- and seventh-grade students. The people in the society presented in the novel have no knowledge of the world outside their community. Only one person in the society, known as "the Giver," holds knowledge of the rest of world. The main character in the novel, Jonas, is designated as the Giver at age 12. In the passage below, Jonas is startled by the look of a newborn baby's eyes. Mr. Goldberg reads the passage aloud to the class:

But he had been startled by the newchild's eyes. Mirrors were rare in the community; they weren't forbidden, but there was no real need of them, and Jonas had simply never bothered to look at himself very often even when he found himself in a location where a mirror existed. Now, seeing the newchild and its expression, he was reminded that the light eyes were not only a rarity but gave the one who had them a certain look—what was it? *Depth*, he decided; as if one were looking into the clear water of the river, down to the bottom, where things might lurk which hadn't been discovered yet. He felt self-conscious, realizing that he, too, had that look.

Which of the following questions would best focus a class discussion on the theme of the passage?

- Why is Jonas startled when he looks into the "newchild's" eyes?
- Why is Jonas chosen to be the Giver?
- Why are there so few mirrors in the community?
- Why has Jonas never bothered to look at himself in a mirror?

### What is this assessment task asking?

This assessment task is asking you to read a short passage from a novel and choose the question that would best focus a class discussion on the theme of the passage. It is not necessary to have read the novel from which the passage is excerpted (Lois Lowry's *The Giver*); however, it is necessary for you to read and comprehend the passage, which includes a number of inferences, and to keep in mind certain information that is presented in the context of this item, including the fact that the character of Jonas was designated as "the Giver," the only person in his society to hold knowledge of the rest of the world. You also need to be familiar with the literary concept of theme and how to identify the theme of a work.

### What information is important?

It is important to read the passage itself closely to identify aspects of the text that might be productive in a discussion of the text's theme. While the main event depicted in the passage is fairly straightforward: Jonas is startled by the look of a baby's eyes, the significance of this event for Jonas is much less straightforward. The explanation of it relies on abstract ideas and figurative language, both of which can be challenging for sixth- and seventh-grade students to interpret. The reference to mirrors could also be a source of confusion. Connecting the rarity of mirrors to the fact that Jonas doesn't regularly think about his own appearance and is therefore startled when the baby's eyes remind him of his own requires a great deal of inference.

It is also important to understand that the character of Jonas has been designated by his society as the Giver, the sole repository of knowledge of the rest of the world. It is also important to know that the theme of a passage or work is the central idea or concept that the author is trying to

communicate. Finally, it is important to not only understand how to identify a theme, but to also to possess knowledge of what kind of questions will help foster student discussion about the theme of a passage.

What is the rationale for selecting an answer?

In order to consider which of the four options would be the best question for generating discussion about the theme of the passage, you must consider the four options in relation to each other.

**Option A:**

The question, “Why is Jonas startled when he looks into the newchild’s eyes?” has the potential to focus the class discussion on the theme of the chosen passage, which is Jonas’ role as the the Giver in a community where the person in this role is the only person with knowledge of the outside world. This question has the potential to elicit from students commentary on a range of issues that relate to this theme. For example, you can imagine students touching on topics such as Jonas’ role as the Giver, his reaction to seeing the newchild’s eyes, and the similarities between Jonas’ own eyes and the newchild’s eyes. This question is likely to generate discussion about possible reasons for Jonas’ reaction to the newchild. Therefore, Option A is a possible answer.

**Option B:**

This question (“Why is Jonas chosen to be the Giver?”) is not the best option because it addresses an overarching facet of the novel (Jonas’s role as the Giver), which is not directly addressed in the chosen passage. In fact, there is no actual mention of Jonas’s role as the Giver in the passage, so any discussion of this question will not be sharply focused on the passage and its theme. While the passage does seem to hint at the idea that Jonas’ own light-colored eyes may have been a factor in his being selected as the Giver, this is not a realization that Jonas reaches in this passage, and discussion of this idea will be more speculation than focused on what is actually present in the given text. Option B is not the best answer.

**Option C:**

This question (“Why are there so few mirrors in the community?”) is not the best option because it addresses a facet of the society that is only tangentially related to Jonas’s reaction to the newchild. While the fact that mirrors are rare helps to create a situation where Jonas is forced to reflect on the idea of light-colored eyes and what their significance may be—without having him do so by looking in a mirror—having students answer this question will foster a discussion that is not focused on the main thrust of the passage, which is *why* Jonas reacts to the newchild having light-colored eyes like his. Option C is not the best answer.



**Option D:**

This question (“Why has Jonas never bothered to look at himself in the mirror?”) is not the best option because, like Option C, it focuses on a facet of the text that is tangential. Since mirrors are rarely used and not made a big deal of in his community, Jonas simply ignores them. Discussion of why he doesn’t look at himself in the mirror will not focus the class on the important information contained in the passage because it can be answered very matter-of-factly and without encouraging much critical thought about the main focus of the passage—Jonas’ reaction to seeing someone else with light-colored eyes like his own and his realization of how light-colored eyes make someone look. Option D is not the best answer.

Option A is the best answer because of its potential to focus students’ discussion on the theme of the passage.

Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Knowledge of the literary concept of theme as the central idea or focus of a passage that the author is trying to communicate.
- Ability to read and comprehend a text passage that includes inferences and figurative language.
- Ability to identify the theme of a text passage.
- Knowledge of the types of questions that have the potential to generate discussion in the service of sixth- and seventh-grade students identifying and comprehending the theme of a text passage.

## Task Design Rationale – Wong: Text Features That Potentially Interfere With Reading Comprehension

Mr. Wong and his fifth-grade students are studying the impact of humans on the environment. The students in Mr. Wong's class are avid readers, often discussing and writing about fiction that they read. Because his students are strong readers, Mr. Wong is surprised that many of them struggle to comprehend informational text. He gives his students the following passage to read in pairs and notes that they can read it fluently.

**Save the Soil!**

Soil seems to be everywhere—in fields, in backyards, clinging to the soles of our shoes. But the nation's fertile soil is vanishing at an alarming rate. For every bushel of corn produced in the U.S., about a bushel of soil disappears.

How? Rain can wash loose soil into streams and rivers. To replace an inch of washed-away topsoil, plants and other matter on the surface must break down for hundreds or even thousands of years. Soil disappears when houses or malls are built on land where crops could be planted. Pollution also ruins soil, making it unsafe for planting.

In the discussion that follows, many students seem to have trouble comprehending the information about soil.

For each text feature, indicate whether it will be likely to cause difficulty.

	Likely to cause difficulty	Not likely to cause difficulty
The connection between "fertile soil" and "bushel of corn produced"		
The synonymous use of "fertile soil" and "topsoil"		
Starting the second paragraph with the one-word sentence "How?"		
The lack of an explicit statement that humans play a role in erosion		
The use of subordinate clauses in a number of sentences		
The description in the first sentence of the text of where the soil might be found		

Excerpt from *TIME for Kids Magazine*, 2/9/1999 © 1999 Time Inc. Used under license. TIME for Kids and Time Inc. are not affiliated with, and do not endorse products or services of Educational Testing Service.

### What is this assessment task asking?

This assessment task is asking you to evaluate different elements of an informational text passage and identify whether each element is either likely or unlikely to cause difficulty for fifth-grade readers. The analysis requires that you pay attention to the fact that in the course of learning about the impact of humans on the environment, the students are reading an informational text that deals with an aspect of this subject (the loss of fertile soil) and are struggling to comprehend it. You must then take into consideration that these students are avid fiction readers who spend a lot of time writing about and discussing the fiction they read. Finally, you need to keep in mind that the students are strong readers who are able to read the passage fluently when assigned to read it in pairs. To respond to this task correctly, you will need to be familiar with the characteristics of fifth-grade readers, including what they are capable of at this level and what they struggle with. For one of the items in the table (“The use of subordinate clauses in a number of sentences”), you will also need to be able to identify this grammatical construct and whether or not it appears in the text passage.

### What information is important?

In responding to this assessment task, you must first identify what the level of reading ability is for the students in this class. In order to evaluate each element listed in the table for whether it is likely or unlikely to cause them difficulty, you must first have a clear sense of their ability. To

that end, you will need to make a note of the fact that these are fifth-grade students who read fiction frequently and who often write about and discuss what they have read. You should be aware that the students are considered strong readers, and the fact that they are struggling to comprehend this informational text passage surprises their teacher. Finally, you should consider that when given the passage to read together in pairs, the students were able to read it fluently.

After considering the overall reading ability of the students, it is important to then consider the features of this particular reading assignment. You need to identify that this passage is an informational text rather than a fictional one. In addition, you need to pay attention to the fact that the students, while able to read the text fluently, are having difficulty comprehending the text.

With this information in mind, you will then need to read the passage with an eye towards identifying any features that might interfere with comprehension for a fifth-grade student with strong reading ability.

What is the rationale for selecting an answer?

**Option A:**

This option should be rated as likely to cause difficulty. For students to fully comprehend the sentence, “For every bushel of corn produced in the U.S., about a bushel of soil disappears,” they must make the connection that corn is grown in soil and that the growing of the corn causes the soil to disappear—a connection that is not explicitly stated in the text.

**Option B:**

This option should be rated as likely to cause difficulty. The text passage does not explicitly state that *fertile soil* and *topsoil* are the same thing, but rather uses these terms interchangeably. While a strong reader would certainly be able to read both of these terms with little problem, this does not guarantee that he or she will understand that they are referring to the same thing. This issue affects passage comprehension and therefore is likely to be an aspect that causes difficulty for the students.

**Option C:**

This option should be rated as unlikely to cause difficulty. The question (“How?”) that begins the second paragraph is answered in the following sentence, “Rain can wash loose soil into streams and rivers,” and elaborated upon in the rest of the paragraph. Since the question clearly asks how soil is lost and is then answered immediately, it would not interfere with comprehension. In addition, you may consider that the students in Mr. Wong’s class are avid readers of fiction and therefore likely to have encountered the literary device of posing a question at some point in their reading.

### **Option D**

This option should be rated as likely to cause difficulty. Humans are not referenced anywhere in the passage, except for an oblique reference in the first sentence (“clinging to the soles of our shoes”). Instead, the passage presents human-created stresses on soil, such as farming, construction, and pollution, but it does not explicitly state that these issues are the result of activities of humans. Students must then make a connection between these stresses and the human cause; failure to make this connection would result in lack of comprehension. Furthermore, you may consider that this reading assignment is being given within the context of studying the impact of humans on the environment and recognize that students may expect a more explicit explanation of the human role in soil erosion and miss the more subtle connections.

### **Option E**

This option should be rated as unlikely to cause difficulty. While there are several subordinate clauses present in the passage, sentence structure is not overly complicated. If the students are strong readers and have experience reading a large amount of fiction, this is a structure they would be quite familiar with. However, to successfully evaluate this option, it is useful to know what a subordinate clause is so that you can first identify whether this construction is even being used in the passage and then identify whether the use of subordinate clauses has the potential to impact student comprehension.

### **Option F**

This option should be rated as unlikely to cause difficulty. The description of where soil might be found is relatively straightforward and moves from the general (“everywhere”) to the specific (“in fields, in backyards, clinging to the soles of our shoes”). Since the students are strong fiction readers, this is a construction that they likely have encountered before in their reading and as a result, it is not likely to impede their comprehension.

### Summary of key knowledge, skills, and reasoning

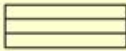

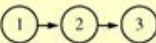
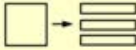
This assessment task draws on the following:

- Knowledge of fifth-grade reading ability and the capabilities of students who are considered strong readers.
- Knowledge of the difference between reading comprehension skills and the ability to read a text fluently.
- Knowledge of the difference in reading comprehension skills required for reading informational texts.
- Knowledge that fifth-grade readers have difficulty recognizing connections that are not explicitly explained in a text.
- Knowledge of subordinate clause structure and knowledge of whether fifth-grade readers encounter this construction frequently in their reading.

## Task Design Rationale – Foreman: Graphic Organizers Supporting Comprehension of Informational Text

Questions 10-11 are based on the following teaching scenario.

Mr. Foreman is teaching his sixth- and seventh-grade students a unit that focuses on strategies for reading informational text. His textbook includes a resource section with ideas that can help students identify the organization of text passages. The following graphic organizers are included in the textbook's resource section:

<b>Description or List</b> 	<b>Compare and Contrast</b> 
<b>Sequence/Time order</b> 	<b>Cause and Effect</b> 

Mr. Foreman is considering how to use these examples to help his students understand how graphic organizers can support their comprehension of informational text.

To best help his students improve their reading of informational text, Mr. Foreman should have them use these graphic organizers to

- record and focus on the meaning of key vocabulary in the text
- comprehend that nonfiction texts are used to extract factual information
- identify concrete visual images to recall important details
- anticipate patterns of information by identifying commonly used structures

### What is this assessment task asking?

This assessment task asks you to identify the best way for students to use a set of graphic organizers to improve their comprehension of informational text. To identify the best choice, you first need to draw on your careful reading of the task or your familiarity with the four graphic organizers to recognize what they have in common. It also helps to know something about either the kinds of tasks sixth and seventh graders usually perform with graphic organizers like these or the level of complexity they can handle when working with informational texts.

### What information is important?

The introduction paragraph in the assessment task provides some critical information about the four graphic organizers pictured: they come from a resource section of a textbook that contains “ideas that can help students identify the organization of text passages.” Drawing on this information and/or on your familiarity with these graphic organizers, you can conclude that they share an important characteristic: they each represent a different way to structure a piece of informational text. This is not a common characteristic for *all* graphic organizers. Some provide students with a structure to support brainstorming; others help them to capture main and supporting ideas from an informational text; still others help them to learn and remember new vocabulary words. However, the four organizers in this assessment task all represent different text structures or organizational patterns.

Now that you know what the four graphic organizers have in common, you can figure out how students can best use them to support their comprehension. First, though, you have to consider the grade level of the students: sixth and seventh grade. By this age, students should have a clear understanding of the difference between informational (nonfictional) text and narrative (usually fictional) text. They should also have had some exposure to the different ways that informational texts can be structured, such as comparison, cause/effect, and problem/solution.

### What is the rationale for selecting an answer?

In order to consider which of the four options would be the best use of the graphic organizers to support improving students' reading comprehension, you must consider each of the four options in relation to the others.

#### **Option A:**

Option A describes a common use of some graphic organizers. While there are specific graphic organizers designed to accomplish this aim, Option A does not describe a plausible use of the four graphic organizers pictured in the assessment task. The introductory paragraph states that the graphic organizers are from a section of the textbook containing “ideas that can help students identify the organization of text passages.” Careful reading of this statement, or your own familiarity with typical uses of the pictured graphic organizers, would lead you to decide that Option A is not the best answer.

#### **Option B:**

While it is true that nonfiction texts can be used to extract factual information, and the four graphic organizers in this assessment task could support students with that process, this use of the graphic organizers would not help students to identify the organization of the text passages they are reading. Furthermore, simply comprehending that nonfiction texts can be used to extract factual information is something that should be very familiar to sixth- and seventh-grade students. For these reasons, Option B is not the best answer.

#### **Option C:**

The graphic organizers pictured here are concrete visual images, and when filled in with information from a text, they might help students to recall important details about that text. However, helping students to recall important details is different from helping them to identify the organization of the text passages. Therefore, Option C is not the best answer.

#### **Option D:**

Each of the four graphic organizers represents a commonly used structure for informational texts. Students who are aware of these four structures can improve their comprehension of informational texts by identifying which of the patterns it follows, if any. If they determine, for example, that the text follows a cause/effect structure, then they can more easily digest the information in the text—and check their own comprehension of it—by understanding that they need to identify some ideas in the text as causes and others as effects. In other words, they can anticipate that the text will link causes to effects, which provides a structure to improve their comprehension. This ability to anticipate patterns of information is a more age-appropriate skill to foster in sixth and seventh graders than the simple awareness, described in Option B, of the

fact that nonfiction texts can be used to extract factual information. Therefore Option D is the best answer.

Option D is the best answer because it explicitly teaches students how to use graphic organizers to support their reading comprehension.

Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Knowledge that graphic organizers must be matched with instructional purpose.
- Knowledge that there is a wide range of uses of graphic organizers, and they cannot be used interchangeably.
- knowledge that description or list, compare and contrast, sequence/time order, and cause and effect represent common text structures found in informational texts.
- Knowledge that sixth- and seventh-grade students typically know the difference between informational (nonfiction) text and narrative (usually fiction) text.
- Knowledge that students who are aware of common text structures can anticipate patterns of information, which supports reading comprehension.

## Task Design Rationale – Able: Responding to Student Writing

Mr. Able's sixth-grade students have just begun writing personal narratives. Manuel produces the following first draft.

Cars are cool. I collect them. They can open the doors & trunk. I can make a list of the cars I know 1st is a Ferari f50, a Viper GTs, orrr a Mustang GT, Lamberginy, corvette, Lexis GT, 1800 Honda GT, Porshe 911 Torbo, Mercedes, Mini. The Mini is cool. It is small and has mini spinners. I imagine where I could ride in my car. I drive in my Mini all day and go to places MC Donald's any time I want. I love car's there just really special to me. There really cool that's wy I like them. If I'm rich I would buy lots of cars. My favorite car is my limo it light's up it's really cool. It reminds me of my to Disney land.

Mr. Able and Manuel have the following conversation during a writing conference.

*Mr. Able:* It seems like you're really interested in collecting cars, that you're really passionate about it. How did you get started?

*Manuel:* I collect cars because they're cool. I was going to bring my cars in for Expert Day.

*Mr. Able:* Let's slow down for a minute. What was the first car you ever collected?

*Manuel:* My mom saw a show on Disney channel about kids who collect things.

*Mr. Able:* Can you tell me about your first car? Do you remember getting it?

*Manuel:* Um... it was my yellow Corvette.

*Mr. Able:* And how did you get your yellow Corvette? Was it a present?

In this writing conference, Mr. Able is attempting to help Manuel improve his narrative by

- focusing his story on a single idea
- developing more vivid descriptions
- including a broader perspective on his topic
- explaining why he likes collecting things

### What is this assessment task asking?

This assessment task is asking you to read both a short draft of a student's essay and a transcript of a writing conference between the student, Manuel, and his teacher, Mr. Able. You are then asked to identify what Mr. Able is trying to accomplish during his conference with Manuel.

### What information is important?

It is important to note that Manuel is in the sixth grade and that his assignment is to write a personal narrative. When reading his narrative, it is important to understand that the main issue with his draft is that while it deals with the topic of collecting cars, it is not focused and includes irrelevant or tangential details that do not support discussion of his hobby. When reading the transcript of Mr. Able's writing conference with Manuel, it is important to focus on Mr. Able's responses and note how he is focusing Manuel's attention. Mr. Able's comments and questions during the conference focus Manuel on identifying or focusing on a topic that could provide structure to the story (e.g., "It seems like you're really interested in collecting cars...How did you get started?")

### What is the rationale for selecting an answer?

In order to consider which of the four options best describes what Mr. Able is doing this conference, you must consider the four options in relation to each other.



**Option A:**

The main problem with Manuel’s draft is that it lacks focus. While there are several issues with spelling and grammar present, these issues are minor when compared to the overarching concept of focus. Manuel begins by stating that he collects cars and then his narrative devolves into a list of facts about cars in general, some of the cars in his collection, or musings on what he would do if owned an actual car, or if he had the money to purchase cars. Furthermore, Mr. Able’s line of questioning during his writing conference with Manuel deals with trying to help Manuel identify a focus for his writing within the general topic of collecting cars. All of the questions that Mr. Able asks Manuel are attempts to get him to think about the first car in his collection (e.g., “What was the first car you ever collected?”). Even when Manuel answers with a tangential or unrelated statement (e.g., “I was going to bring my cars in for Expert Day”), Mr. Able responds with another question focusing on how Manuel started his collection, specifically focusing on the first car he acquired. From this line of questioning, it is apparent that Mr. Able is attempting to help Manuel identify a specific instance—the way he began collecting cars and the first car he ever acquired—that can become the focus of his narrative. Therefore, Option A is likely the best answer.

**Option B:**

Manuel’s draft does include two brief descriptions of specific cars, the Mini and the limo, although these descriptions are not especially vivid. However, improving these descriptions or developing new ones would not be the next logical step in revising his draft. Once he narrows his focus, he can start adding description to the topic, but first he needs a topic. Mr. Able’s question, “Can you tell me about your first car?” might be interpreted as a request for a description, but the fact that it is immediately followed by the question, “Do you remember getting it?” and later, the follow-up, “And how did you get your yellow Corvette?” suggest that his primary goal is not to help Manuel to develop vivid descriptions. Therefore, Option B is not the best answer.

**Option C:**

Manuel’s draft does not suffer from the lack of a broader perspective; in fact, the problem is the opposite of this issue—a lack of focus. Furthermore, Mr. Able’s questions during their writing conference attempt to get Manuel to focus on a single, very specific idea—the first car he acquired for his collection—and therefore, these questions would not help Manuel to identify a broader perspective, but rather would focus his writing on a smaller idea. Therefore, Option C is not the best answer.

**Option D:**

Option D (“explaining why he likes collecting things”) is not the best answer for this question. First, in his draft, Manuel gives some indication of why he enjoys collecting cars (e.g., “Cars are cool”). In addition, during their writing conference, Mr. Able never asks Manuel to explain why

he enjoys collecting cars (or anything else for that matter). In fact, Mr. Able ignores Manuel's comment about a television show that features kids who collect things ("My mom saw a show on Disney Channel about kids who collect things"), which would have been a good lead-in for Mr. Able to then ask Manuel about why he thinks kids enjoy collecting things and why Manuel, in particular, likes collecting cars. Instead, all of Mr. Able's questions focus on how Manuel got started collecting cars and, even more specifically, on the first car he acquired for his collection.

Option A is the best answer, because in the writing conference, Mr. Able is attempting to focus Manuel on a single idea: getting his first toy car, a yellow Corvette.

#### Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Knowledge of personal narrative and the characteristics of the genre.
- Knowledge of the stages of the writing process, particularly as they relate to narrative writing.
- Ability to recognize a teacher's use of questions to help a student develop a more focused topic.
- Ability to guide students with probing questions related to a specific instructional goal.
- Ability to identify strengths and weaknesses in student writing.

## Task Design Rationale – Okeke: Showing Versus Telling

Ms. Okeke's fifth-grade class is reading a short story entitled "The Terrible Summer." She wants to use sentences from this story to model how the author uses the technique of "showing" rather than "telling" in her writing.

Which of the following sentences should Ms. Okeke select to best demonstrate the concept of "showing"?

- The sun flamed like a giant match, and the mercury was hovering at a hundred degrees.
- It was such a feverish afternoon, everyone felt like they were in an oven.
- The asphalt bubbled, and ice cream cones melted before the fifth lick.
- Bill detested the sweltering summer, so he conspired to stay in the delicious coolness of his basement.

What is this assessment task asking?

This assessment task asks you to choose the example that is the best model to use with fifth graders to model showing rather than telling.

What information is important?

It is important to notice that the assessment task asks for an example of *showing*, not an example of several other characteristics often associated with effective writing, such as sophisticated vocabulary, figurative language, or alliteration. It is also important to understand the difference between showing (indirectly conveying an idea through the use of specific details or images) and telling (directly describing what you want the reader to know or think). Finally, you should bear in mind that clichés can interfere with the effectiveness of a literary technique; original, imaginative images and details are much more effective at showing than tired or overused images or details.

What is the rationale for selecting an answer?

In order to consider which of the four options is the best model of showing, you must consider the four options in relation to each other.

### **Option A:**

This sentence includes a vivid simile (“the sun flamed like a giant match”) and a clause that could be described as showing how hot it was (“the mercury was hovering at a hundred degrees”), rather than telling. However, the reference to mercury is a fairly commonplace way of describing the temperature, so the second clause might be considered a cliché or at least an unimaginative example of showing. Option A is not a strong example of showing.

**Option B:**

This sentence also includes a simile (“everyone felt like they were in an oven”), but the simile is a bit overused, and the first clause in the sentence (“It was such a feverish afternoon”) is a clear example of telling, even though it includes the sophisticated vocabulary word, *feverish*. Option B is not a strong example of showing.

**Option C:**

This sentence is a clear example of showing. Rather than using direct or figurative language to describe the heat, it demonstrates the heat indirectly by describing the impact that it had on asphalt and ice cream cones. Furthermore, the images of bubbling asphalt and quickly melting ice cream, while familiar enough to people who have been in heat waves, are not clichéd references. Option C is a strong example of showing.

**Option D:**

This sentence uses sophisticated vocabulary (*detested, conspired*) and alliteration (“sweltering summer”), but it still tells rather than shows. We are told that Bill hated the hot summer and therefore stayed in his basement. Option D is not a strong example of showing.

Option C is the best answer.

**Summary of key knowledge, skills, and reasoning**

This assessment task draws on the following:

- Knowledge that the literary technique of showing is indirectly conveying an idea through the use of specific details or images. It is the opposite of telling, directly describing what you want the reader to know or think.
- Knowledge that sophisticated vocabulary, figurative language, and alliteration may be present in a literary text that tells rather than shows.
- Knowledge that showing may not be as effective when descriptions utilize tired or overused images or details.

## Task Design Rationale – Figueroa: Helping Students Write Similes

Ms. Figueroa and her students are reading an excerpt from the Wallace Stevens poem "Anecdote of the Prince of Peacocks."

I knew the dread  
Of the bushy plain,  
And the beauty  
Of the moonlight  
Falling there,  
Falling  
As sleep falls  
In the innocent air.

After discussing the definition of simile, Ms. Figueroa asks students to write down as many similes as they can think of that describe moonlight. She notices that Regine, a struggling student, isn't writing at all. When Ms. Figueroa asks her why, Regine says, "Writing a simile about moonlight doesn't make sense. Moonlight is just like moonlight; it isn't like anything else."

Which of the following teaching responses is most likely to help Regine write a simile for moonlight?

- "Of course it isn't *exactly* like anything else. What you need to do is think about a comparison."
- "What color is the moonlight? Can you think of something else that has the same color?"
- "The sunlight is a lot like moonlight, isn't it? Except sunlight is warm and moonlight isn't."
- "Remember last week when we talked about how the sun is a flame? We are doing something similar here."

"Anecdote of the Prince of Peacocks" from *THE COLLECTED POEMS OF WALLACE STEVENS* by Wallace Stevens, copyright 1954 by Wallace Stevens and renewed 1982 by Holly Stevens. Used by permission of Alfred A. Knopf, a division of Random House, Inc.

### What is this assessment task asking?

This assessment task asks you to choose the best response to help a student who is unable to write a simile. To respond correctly, you need to first assess the nature of the student's difficulty. Reading the poem excerpt beforehand is not essential but may be helpful in forming a clearer sense of why Regine is confused. The simile in the poem excerpt is very abstract, and Regine likely needs a more concrete example to help her understand what a simile is. Based on Regine's statement, you can conclude that, despite hearing her teacher's definition of simile, she doesn't understand that writing a simile requires finding a common thread that links two otherwise unlike things. Next, you must determine which of the four proposed responses to Regine would best help her to write a simile. To choose the best response, you will find it helpful to draw on your knowledge of how similes and metaphors function, how to make complex writing tasks more accessible to students, and the importance of introducing abstract concepts with concrete, familiar examples.

### What information is important?

It is important to notice that Ms. Figueroa has already discussed the definition of simile with the class. After reading the poem, you might infer that she has also pointed out to her class that the poem excerpt contains a simile comparing the way moonlight falls on a plain to the way sleep falls in the air. You might notice that this is a fairly abstract simile, a very challenging example to use when first introducing the concept. Regine's response is important because it provides a window into the nature of her difficulty. "Moonlight is just like moonlight; it isn't like anything else" suggests that she doesn't understand that writing similes requires finding one commonality between two otherwise unlike things.

In summary, we know the following about Regine:

- She has heard her teacher’s definition of a simile.
- She may have read the excerpt, which includes a very sophisticated simile.
- She does not seem able to write a simile comparing moonlight to anything.
- She does not seem to understand that writing a simile requires finding a common thread that links two otherwise unlike things.

What is the rationale for selecting an answer?

In order to consider which of the four teaching responses is most likely to help Regine, you must consider the four options in relation to each other.

**Option A:**

You may decide to reject this response right away, because the fact that it starts with “of course” could be read as giving it a condescending or disrespectful tone that does not seem appropriate to use with a student—especially a struggling student. It is also not the best answer because it doesn’t provide Regine with any new information about writing similes. If Ms. Figueroa has already defined a simile for the class, then she must already have explained that it is a comparison of two things that aren’t alike. Option A is not a strong choice.

**Option B:**

This response scaffolds the process of writing a simile for Regine by choosing a characteristic for her that might serve as the common thread, linking moonlight to something else that is otherwise different. In this case, the teacher chooses color as the common thread, which is a much more concrete characteristic than the one used in the poem excerpt: the way moonlight falls on a plain. The choice of a concrete characteristic is important; you can infer that Regine does not understand the meaning of the abstract simile in the poem, so she needs to start with a concept that is easier to grasp. By providing the characteristic of color, Ms. Figueroa is essentially breaking Regine’s simile writing task into three steps and completing the first step for her. Regine now needs to choose another object that has the same color as moonlight, and write a sentence that uses *like* or *as* to link it to moonlight. Option B is a strong choice.

**Option C:**

In this response, the teacher chooses another concept for Regine to compare to moonlight: sunlight. This is not the most helpful choice, because sunlight and moonlight are actually pretty similar. Drawing on your knowledge that powerful similes are comparisons of essentially unlike things, you would reject this option because it is not likely to guide Regine toward an effective simile for moonlight. Option C is not a strong choice.

**Option D:**

In this response, the teacher tries to help Regine write a simile by reminding her about a previous class discussion about metaphor. In general, helping students to connect new learning to background knowledge gained from previous lessons is a good idea. In this case, however, the reference may remind Regine that there is a similar form of figurative language called metaphor, but it does not give her the practical help she needs to write a simile. Furthermore, Regine's earlier statement, "Moonlight is just like moonlight; it isn't like anything else" suggests that she probably understood metaphors no better than similes, since both are comparisons between unlike things.

Option B is the best answer because it provides Regine with the direct, scaffolded support to write a simile.

Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Knowledge that a simile is a comparison of two things that have a common characteristic but otherwise are unlike. It typically uses *like* or *as* to compare the two things.
- Knowledge that when a student is struggling with a complex writing activity, it is helpful to break the activity into separate steps and sometimes to complete the first step for or with the student.
- Knowledge that physical characteristics such as color are concrete and thus provide a useful first step into constructing similes.
- Awareness that, if a student is struggling with the topic of simile, introducing the topic with a concrete simile rather than an abstract simile might support the student's development of knowledge of similes.

## Task Design Rationale – Jackson: Identifying Strengths in Student Writing

Mr. Jackson's sixth- and seventh-grade students are writing flash fiction. After reading the following draft of Arielle's short story, Mr. Jackson wants to point out some of the strengths in her writing.

Natasha left the house and walked sadly down the sidewalk, slowly shuffling her feet. As she walked by the old house, she was surprised by how sad she felt. She stepped onto the grass, and the sprinklers sprayed her face. The sprinkle of cold water caught her by surprise. Her brother shouted down the street asking her to wait up. Thankfully he couldn't tell the tears from the drops of water.

For each writing feature, indicate whether it describes a strength in Arielle's writing.

	<b>Describes a strength in Arielle's writing</b>	<b>Does not describe a strength in Arielle's writing</b>
Strong character development		
Use of figurative language		
Conveying emotions through showing		
Use of powerful images		

### What is this assessment task asking?

This assessment task is asking you to evaluate a short piece of fiction written by a student. The task presents four features of fiction writing, and you must determine whether or not each of them is a strength present in the student's writing. This evaluation requires you to consider that the student, Arielle, is in either sixth or seventh grade and that the class is writing fiction. It is not necessary to be familiar with the genre of flash fiction to successfully answer this task; however, it is necessary for you to be able to identify what constitutes character development and the use of figurative language. It is also helpful to know what is possible in short pieces of writing to provide a basis for evaluating the student writing. In addition, you should be familiar with the concept of "showing rather than telling" and be able to identify the use of powerful images in writing.

### What information is important?

It is important to understand that the class is writing fiction and that, as a sixth or seventh grader, Arielle is unlikely to be a sophisticated fiction writer. In other words, you need to evaluate each of the writing features using appropriate standards for this age. It is also important to have an understanding of what strong character development is and be able to identify it in a short piece of writing. In addition, knowledge of the elements of figurative language, such as simile and metaphor; how figurative language differs from simple description; and the ability to identify figurative language in writing are also essential. Finally, an understanding of how a writer shows emotion rather than tells about it is also essential to completing this assessment task.



### What is the rationale for selecting an answer?

Once you have read through the excerpt, begin reading through the writing features in the table and for each option decide whether it describes a strength.

#### **Row A:**

This feature (“Strong character development”) is not a strength of Arielle’s writing as demonstrated in the given story. The story does not provide any specific information about Arielle’s character aside from the basic information that she has a brother and is sad. As such, it is impossible to form any concrete judgment about her character from the story, meaning that character development is not present in the story. The best answer for this row is “Does not describe a strength in Arielle’s writing.”

#### **Row B:**

This feature (“Use of figurative language”) is not a strength of Arielle’s writing. Arielle does not employ the most common types of figurative language—simile, metaphor, or personification—or any of the less common types, such as hyperbole or oxymoron. Therefore, the best answer for this row is “Does not describe a strength in Arielle’s writing.”

#### **Row C:**

This feature (“Conveying emotions through showing”) is a strength of Arielle’s writing. Although Arielle directly states that Natasha “was surprised by how sad she felt,” she also demonstrates Natasha’s state of mind by describing how she is moving (“slowly shuffling her feet”) and by providing the detail that Natasha is thankful her brother “couldn’t tell the tears from the drops of water” on her face. From these details a reader could infer that Natasha is sad even without being directly told this fact by the author. The comment about shuffling feet might be considered a cliché in adult writing, but for a sixth or seventh grader, it suggests a reasonable level of skill at showing. Therefore, the best answer for this row is “Describes a strength in Arielle’s writing.”

#### **Row D:**

This feature (“Use of powerful images”) is a strength of Arielle’s writing. There are several vivid images in this story: Natasha slowly shuffling down the sidewalk, the sprinklers spraying cold water, and her brother calling after her down the street. The final image of Natasha’s tears mixing with the drops of water from the sprinkler is particularly compelling and vivid. Therefore, the use of powerful images is a strength of Arielle’s writing as demonstrated in this story. The best answer for this row is “Describes a strength in Arielle’s writing.”

### Summary of key knowledge, skills, and reasoning

This assessment task draws on the following:

- Knowledge of what strong character development is and the ability to identify it in a piece of writing.
- Knowledge that figurative language consists of the use of literary devices such as simile and metaphor and the difference between these devices and simple description.
- Knowledge of techniques used by authors to show emotion rather than simply telling about it.
- Ability to identify powerful images in a text.
- Familiarity with the level of fiction writing typical for sixth and seventh graders.