Are Multiple Measures Meaningful?:
Lessons From a Statewide Performance Assessment

Gail Lynn Goldberg

Educational Consultant
Charles County Public Schools, Maryland

Barbara Sherr Roswell

Department of English
Goucher College

For practical, psychometric, and pedagogical reasons, strong interest exists in developing multiple-measure constructed-response items for use in large-scale performance assessments. Items that can be scored for evidence of proficiency in 2 or more content areas raise questions, however, about the “fit” between various content areas and the possibility of sending cross-messages or confounding different content demands. To determine the factors that contribute to or compromise the effectiveness of multiscored items, in this study we combine analysis of statewide score data from the 1996 Maryland School Performance Assessment Program tests, administered at Grades 3, 5, and 8, with systematic analysis of 60 activities providing measures of writing, language usage (LU), or both, as well as one or more content areas. Although test developers to date have had greater success in creating writing/LU items that can also be scored for reading and social studies than for mathematics and science, we argue for the validity of multiple-measure items across all content areas and suggest that, across content areas, successful multiple-measure items (a) make information sources explicit and allow students to draw on both text-based and personal knowledge; (b) identify the specific content area skills or concepts being assessed; (c) permit open-ended development; (d) maintain a good fit between content demands and the rhetorical situation, creating an authentic writer–audience relationship; and (e) are uncluttered, focused, and direct, with good recall capability. Thus, we suggest that
test developers reorient their concerns from the difficulty of items to identifying elements of multiple measure activities that facilitate or impede students’ ability to demonstrate what they know and can do in different content areas.

Today, those who are interested in the development of performance assessments have available an ever-increasing number of resources to guide them. Many teachers and test developers begin by immersing themselves in the abundant research on the design of writing prompts, one of the most common types of performance tasks (e.g., Brossell, 1983; Greenberg, 1981; Hoetker, 1982; Ruth & Murphy, 1988). They are also likely to consult design standards formulated in recent years to guide the classroom implementation of performance-based instruction and assessment (Stiggins, 1996; Wiggins, 1992, 1998; Wiggins & McTighe, 1998). Such standards or guidelines concur that good performance tasks will be engaging, integrative, and authentic; will allow for open-ended responses; and will “elicit the right student performance while avoiding things that might interfere with [that] performance” (Arter, 1996, V–1: 2). As useful as these sources are, they have certain limitations, particularly when it comes to the design and implementation of large-scale standardized tests bearing the appellation “performance assessment.” Instead of single prompts or more holistic tasks, these tests are comprised of a series of constructed-response items yielding products or performances such as writing, drawing, and displays of data. Although knowledge about crafting effective writing prompts and understanding of general principles of performance task design may inform the development of constructed-response items, there is to date little information available specifically on the item level to guide effective test development within a performance assessment context (Goldberg, 1996; Shavelson & Baxter, 1992). And, although these few sources may help in the creation of constructed-response items measuring proficiency in a single content area, such as science or social studies, we are on virtually uncharted ground when developing items intended to provide multiple measures—items that each yield a response that may be examined for evidence of what students know and can do in more than one discipline.

Strong interest exists among the assessment community in the opportunity to obtain multiple measures from a single constructed-response item, effectively getting “two for the price of one.” Under the sponsorship of the Council of Chief State School Officers, a subgroup of the 21-state collaborative on assessment and student standards has been working on technical guidelines that address this option among other aspects of performance assessment (Winter, personal communication, January 21, 1997). The National Assessment of Educational Progress has also recently commissioned studies on multiscoreing (Atash, White, & Kruglanski, 1997). The direction of this research has thus far focused on procedural issues (e.g., how multiple measures might be obtained within existing assessment structures), and has yet to address the complex questions associated with developing performance assessments designed specifically to yield multiple measures.
The possibility of drawing performance data in different content areas from the same responses appeals to the assessment community for a number of reasons. Certainly, there are practical implications including reduced time and cost for item development, test administration, answer document processing, and handscoring. Furthermore, efficient use of classroom time for testing helps to make any instrument more student- and teacher-friendly. From a psychometric perspective, the possibility of using information from multiple-measure items to augment data obtained from single content area assessments is attractive. Finally, this option appeals not only for practical and psychometric reasons, but also pedagogical reasons. The use of interdisciplinary knowledge in real-life thinking and problem solving underlies many of the newer curricular frameworks (Ackerman & Perkins, 1989; Drake, 1993; Jacobs, 1989), and scoring students’ work for evidence of proficiency in multiple content areas promises to mirror best practice and support content area integration in instruction. If we believe that students should be assessed in much the same ways as they learn, then assessment activities that integrate multiple disciplines are desirable (Herman, Aschbacher, & Winters, 1992). Nevertheless, using such items inevitably raises questions about the quality of information we can obtain from activities that may confound content and rhetorical demands or send other sorts of cross-messages. We may wonder whether a student’s lack of mastery of a particular skill (e.g., in mathematics) may mask the ability to write cogently, and conversely, whether the development of a compelling argument, for example, may divert a student’s attention from the content area concepts being measured and distort evidence of understanding a given principle.

The purpose of this study, therefore, is to examine a variety of questions related to the measurement of writing and language usage (LU) within the context of the Maryland School Performance Assessment Program (MSPAP), a statewide, content-based performance assessment of all students at Grades 3, 5, and 8. Designed as part of a larger school improvement initiative, MSPAP has, since its inception in 1991, included multiple-measure items among the variety of constructed-response item types that include writing, drawing, development of graphic organizers, and data displays. The use of items measuring both content area knowledge and writing or LU was the natural outgrowth of the fact that many of the test items across content areas require students to construct written responses at least one paragraph in length and that the task format itself provides needed motive and material to support the generation of text. In its design, MSPAP is representative of other large-scale performance assessment tests, and thus, the problems and promises of multiple measures it illustrates have far-reaching implications for the development, implementation, and scoring of other assessments at the local, state, and national level.

In exploring the potential of multiple-measure items, we wished to determine in particular (a) whether performance in writing and LU differs (and if so, how?) based on linkage to the various content areas (reading, mathematics, science, and social studies) and how performance in those areas in turn may be affected by
items having to do “double or triple duty”; (b) what particular features of constructed-response activities facilitate or impede the opportunity to obtain accurate information about student performance in multiple areas; and (c) what implications such findings ought to have for test design. Based on the practice in MSPAP of obtaining measures of writing and/or LU as well as one or more content areas including reading, mathematics, science, and social studies, the overriding question is not whether items can be scored for multiple content areas, but how they need to be designed so that this goal can be achieved in a valid and reliable way.

BACKGROUND

Typically, the direct assessment of writing involves applying scoring criteria to a student response and obtaining a measure of writing ability alone. Similarly, most performance assessment tasks in other content areas are intended to provide a measure only of that area. However, when students respond to activities in various content areas by producing written text, those responses may be scored sequentially—that is, by applying a series of different evaluative criteria to measure distinct learning objectives or outcomes (Goldberg, 1993; Goldberg & Michaels, 1995). This sequential approach is intended to ensure that decisions about writing scores are not confounded by evidence of content knowledge, or conversely, that decisions about content area proficiency are not confounded by skill in mechanics or sentence formation. MSPAP embeds writing and LU measures within content areas activities in thematically coherent performance assessment tasks and, therefore, routinely utilizes this sequential approach to scoring.

Although many assessments subsume issues of correctness within a larger writing construct, on MSPAP, separate scale scores are determined for writing (focusing on such dimensions as organization, development of ideas, and attention to audience) and LU. This distinct Maryland Learning Outcome (MLO) measures not only correctness in usage, mechanics, and sentence formation, but also (according to the LU scoring rubric) “word and sentence order and language choices to express meaning with style and tone.” As such, LU encompasses and treats separately some of what are considered additional “dimensions” of writing in a number of other assessment programs. The area of writing, in turn, includes three different outcomes—informative, persuasive, and expressive writing—matrix-sampled through three test forms and each scored with purpose-specific criteria. With the occasional exception of items measuring expressive writing (stories, poems, and plays), no brief constructed-response measures of writing and LU are freestanding, but occur instead through sequential scoring of content-based items. The placement of a familiar writing or LU icon next to an activity in the student answer booklet alerts students that the response to that item will be scored not only for a particular content area, but for writing or LU as well.
In typical task design, although LU items occur at varying points in different tasks, LU/writing items are usually culminating activities. In tasks that contain a writing item, the introduction and preassessment activities typically alert students that they will be gathering information and organizing ideas toward the goal of presenting in writing, for an appropriate audience, their perspective or solution to a problem. In all writing and most LU items, full rhetorical specifications (topic, audience, purpose, and form) are given to add authenticity to the writing activity and to mirror recommended instructional practice. All writing items also include scaffolding that reminds students of the characteristics of texts written for the stipulated purpose—to inform, to persuade, or to express personal ideas.

MSPAP tasks are all developed by teachers with expertise in the particular content areas and grade level for which the task is intended. Typically, teams are given a general topic, a suggested context, and an assignment of measures to be obtained over the course of multiple activities within the task. Content measures each address one or more of the MLOs and their subsets or indicators—the specific skills, processes, and understandings expected of students by the end of Grades 3, 5, and 8. Most often, activities first take form as content measures, with cuing then overlaid to address the need for the item to serve as a writing and/or LU measure as well. This process generates some predictable tensions between the essential demands of each of the areas being measured, and care is taken during the revision and refinement process to address these tensions.

Besides developing a coherent set of activities, task-writing teams are responsible for drafting the evaluative criteria by which content area scores will be assigned. Although the scoring tools for each of the purposes for writing and for LU are generic (i.e., they are not modified to fit specific items), all content scores are assigned through the application of activity-specific scoring tools. Each of these scoring tools is comprised of a stem statement that identifies at outcome and indicator level the specific skills, processes, or knowledge that the activity is intended to assess; the stem is followed by a set of descriptors to differentiate levels of performance. Score scales for MSPAP items vary (typically from 0–1 to 0–3) based on the complexity of the item. The descriptors for Score Point 1 always define performance that is at least satisfactory, whereas those for the highest score point define excellence. In crafting activity-specific scoring tools, task writers are expected to conform to various guidelines, the foremost being to score only for that which the activity makes clear that the students need to do.

MSPAP is scored by Maryland teachers assigned to teams who score a portion of the items in each booklet. Typically, all writing and LU measures are scored by one team for each test form. Therefore, when an item measures both writing and LU, the same scorer will apply different evaluative criteria to the same response. A member of a different team applies yet another scoring tool to obtain a content measure for that same item. Because MSPAP is intended to serve as a school level accountability and instructional improvement program, and is not intended to
yield individual student scores, items are not double scored for the same outcome measure. Because interrater reliability cannot be determined, as a proxy for that agreement data, scorers are given periodic validity and check sets, and agreement rates are determined by comparison with preestablished “true scores.” In 1996, the overall agreement rate across grade levels and content areas was about 86% and was not significantly different for writing than for all other areas but mathematics, which tended to be slightly higher (Measurement Inc., 1996).

METHODS AND DATA SOURCES

The nature of performance assessment lends itself to investigation that is multimethodological and integrates data from a rich variety of sources. Long-standing experience with MSPAP and our familiarity with the literature on performance assessment contributed to our skepticism about generalizing about all multiscored items or treating them from a primarily statistical perspective as traditional item analysis has often done. Believing that in both the design and the evaluation of performance assessment, “the devil is in the details,” in this study we combined the close and systematic textual analysis of test items with statistical data on student performance on these activities.

Development of Descriptive Item Profiles

We began by creating a detailed, descriptive analysis of all multiple-measure LU and writing items (N = 60) from two of three 1996 MSPAP test forms per grade level, each taken by approximately 20,000 students. We did so to examine the possible impact of each item’s language and format on student performance within and across content areas. For each item, we developed a brief synopsis and recorded basic features: general topic, position during the testing period (day and task order), position within task (beginning, middle, or end), number of lines allocated to the response in the answer booklet, content areas scored, and rhetorical specifications of audience (familiar or unfamiliar, peer or adult) and response form (report, note, summary, etc.). See Figure 1 for a sample descriptive item profile.

The heart of any activity or prompt are the cues it uses to direct students regarding precisely what they need to do, about what, and drawing on what bodies of knowledge. Therefore, we next codified all directives under the broad categories of (a) content and action cues, (b) development cues, and (c) cues for information sources. Under content and action cues, we identified the number and range of content topics (specific things to talk about, sometimes broken into subtopics) and cuing verbs (words like explain, describe, or tell, signaling what students actually had to do). We identified development cues as either open-ended (encouraging
considerable choice of content, approach, or both) or restricted (limiting the range of information possible to extend and support the ideas in the response). Under cues for information sources, we differentiated between explicit and implied sources (“Using what you know and have learned in this task” vs. “You will now”). Within those categories, we further differentiated between general versus specific information source cues (“Using what you’ve read” vs. “Using information from the chart titled ‘Soundwaves’”). We also distinguished between test-based sources (charts, pictures, readings, prior task activities, etc.) and prior knowledge, which we categorized as either content-based (like a mathematics formula) or personal (perspectives acquired through experience).

Statistical Analysis and Identification of Focus Items

The second stage of our investigation incorporated analyses of statistical data from these same test forms, including the following for each item: scale score location, average raw scores, average as percentage maximum, and score point distributions. All MSPAP items are scaled using a “two parameter partial credit”
model, a generalization of Masters’s Partial Credit model (Yen & Ferrara, 1997), to meet various program requirements such as the establishment of proficiency levels and outcome scores comparable across test forms over time. Among item statistics derived from this process are the item’s scale score location for each content area, indicative of the difficulty of the item and used to determine the proficiency levels reported for accountability purposes. For MSPAP, the score scale ranges from 350 to 700 and has a mean of about 500 and a standard deviation of about 50 (CTB Macmillian/McGraw Hill, 1992). For each content area, we then determined the mean scale score locations in each test form for (a) all items measuring that content area (both with and without LU), (b) all items measuring content only (without LU), and (c) all items measuring content and LU. Next, we compared scale score locations for all multiple-measure items with other items measuring the same content area, to identify outliers and, hence, potentially problematic items. We considered as outliers any items whose scale score location differed from the mean scale score location for that content area–test form by more than 20 (equivalent to the mean deviation in many of the content areas within the test forms we investigated). We did not assume that outliers were necessarily problematic; indeed, we would expect that in a well-designed assessment, some items will be significantly harder or easier than others. Rather, we flagged outliers as focus items meriting scrutiny to determine whether their design as multiple-measure items might be distorting evidence of proficiency in any or all of the areas being assessed.

Although we used scale score locations to identify focus items, we returned to raw scores to conduct a completion analysis that allowed us to distinguish nonattempts (coded as blank responses but assigned a score of zero) from true zero scores (indicating attempted but not yet satisfactory performance). This information allowed us to consider such issues as the placement and timing of tasks and activities as possible causes of nonattempts and, therefore, to refine our interpretations of both scale score data and the descriptive analyses of items.

Data Synthesis

Having identified outliers, we returned to our descriptive profiles to engage in closer analysis of the characteristics of the focus items in relation to the other multiple-measure items and to determine which, if any, patterns in cuing and format emerged that might account for the flagged items’ atypical item scale score location. At this stage, our analyses were also informed by extensive reading of sample responses to each of these items. This comprehensive analysis enabled us to generalize about features of items that effectively allow students to show what they know and can do in multiple content areas and about features of items that may distort the measurement of students’ proficiency in one or more content areas.
RESULTS AND DISCUSSION

Across content areas and grade levels, comparison of scale score locations for content-only and multiple-measure items should allay concern that the inclusion of multiple-measure items distorts information about students’ proficiency in a given content area. Given the test design with its limited number of multiple-measure items per test form, in no content area across grades is the difference between mean scale score location of content-only items and mean scale score location of all items (content only and multiple measure) significant (see Table 1).

Nevertheless, in some content areas, a difference between mean scale score locations for items with and without linkage to LU is evident. Examination of item scale score locations within and across content areas for each form and grade level seems to suggest that combining science or mathematics content (reported on MSPAP separately from mathematics process) with LU may yield activities that are more difficult than those providing a measure of these content areas alone. Specifically, although single and multiple-measure items yielded fairly comparable content area item scale score locations in reading, mathematics process, and social studies, with 14 out of 42 items (33%) flagged, a differential in location of 20 or more was noted in 17 of 26 items (65%) measuring LU and mathematics content or science. This might lead one to believe that adding rhetorical demands will sometimes interfere with or mask evidence of students’ proficiency in certain content areas.

Measures of LU showed less variation that could be attributed to linkage with different content areas (see Table 2). Only in two of six test forms, and only in mathematics content and mathematics process, did mean scale score locations for items measuring LU plus a content area differ from the mean scale score location for all LU items by more than 20. These results indicate that, with the possible exception of a small number of mathematics items, designing items to measure LU in addition to one or more content areas does not introduce distortions into the measurement of proficiency in LU.

When we shift attention from scale score data in isolation to analysis of score data in conjunction with descriptive item profiles, we find that features other than content area alone account for the way an activity functions as a source of multiple measures. That is, among the various items providing multiple measures in math content and science were several that defied the pattern of greater difficulty, whereas there were several outliers in the other content areas that raised questions about possible distortion associated with linkage to LU. Thus, although some content areas provided more instances of outliers—items that might be failing to provide accurate evidence of what students know and can do in a specific content area—we found that this phenomenon was more likely attributable to specific and predictable features of the individual activities than to the nature of the content area per se. As the illustrations and discussion that follow demonstrate, coupling
### TABLE 1
Comparison of Mean Scale Score Locations for All Items, Content-Only Items, and Items Measuring Content and Language Usage by Content Area

<table>
<thead>
<tr>
<th>Grade</th>
<th>Form</th>
<th>All Items N</th>
<th>Mdev</th>
<th>Content Area Only Items N</th>
<th>Content Area and Language Usage Items N</th>
<th>Individual Content Area and Language Usage Items N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reading</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>E</td>
<td>546</td>
<td>13</td>
<td>21</td>
<td>542</td>
<td>564</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>531</td>
<td>12</td>
<td>23</td>
<td>529</td>
<td>549</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>520</td>
<td>30</td>
<td>27</td>
<td>515</td>
<td>544</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>540</td>
<td>12</td>
<td>27</td>
<td>544</td>
<td>523</td>
</tr>
<tr>
<td>8</td>
<td>E</td>
<td>521</td>
<td>14</td>
<td>20</td>
<td>518</td>
<td>540</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
<td>531</td>
<td>30</td>
<td>20</td>
<td>530</td>
<td>534</td>
</tr>
<tr>
<td><strong>Math content</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>E</td>
<td>536</td>
<td>30</td>
<td>31</td>
<td>533</td>
<td>570</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>526</td>
<td>25</td>
<td>36</td>
<td>526</td>
<td>550</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>542</td>
<td>25</td>
<td>43</td>
<td>540</td>
<td>590</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>533</td>
<td>26</td>
<td>49</td>
<td>535</td>
<td>589</td>
</tr>
<tr>
<td>8</td>
<td>E</td>
<td>560</td>
<td>25</td>
<td>25</td>
<td>558</td>
<td>577</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
<td>546</td>
<td>33</td>
<td>30</td>
<td>546</td>
<td>NA</td>
</tr>
<tr>
<td><strong>Math process</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>E</td>
<td>544</td>
<td>14</td>
<td>15</td>
<td>541</td>
<td>557</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>539</td>
<td>9</td>
<td>27</td>
<td>526</td>
<td>550</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>567</td>
<td>14</td>
<td>19</td>
<td>567</td>
<td>568</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>558</td>
<td>12</td>
<td>35</td>
<td>551</td>
<td>575</td>
</tr>
<tr>
<td>8</td>
<td>E</td>
<td>562</td>
<td>11</td>
<td>17</td>
<td>563</td>
<td>558</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
<td>531</td>
<td>7</td>
<td>57</td>
<td>512</td>
<td>559</td>
</tr>
<tr>
<td><strong>Science</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>E</td>
<td>512</td>
<td>19</td>
<td>43</td>
<td>505</td>
<td>547</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>518</td>
<td>18</td>
<td>45</td>
<td>506</td>
<td>557</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>546</td>
<td>26</td>
<td>50</td>
<td>539</td>
<td>597</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>529</td>
<td>23</td>
<td>32</td>
<td>528</td>
<td>534</td>
</tr>
<tr>
<td>8</td>
<td>E</td>
<td>544</td>
<td>25</td>
<td>31</td>
<td>541</td>
<td>568</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
<td>560</td>
<td>21</td>
<td>24</td>
<td>559</td>
<td>570</td>
</tr>
<tr>
<td><strong>Social studies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>E</td>
<td>521</td>
<td>19</td>
<td>21</td>
<td>521</td>
<td>521</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>516</td>
<td>20</td>
<td>29</td>
<td>513</td>
<td>530</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>560</td>
<td>17</td>
<td>38</td>
<td>553</td>
<td>562</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>560</td>
<td>20</td>
<td>30</td>
<td>560</td>
<td>555</td>
</tr>
<tr>
<td>8</td>
<td>E</td>
<td>534</td>
<td>19</td>
<td>31</td>
<td>533</td>
<td>541</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
<td>539</td>
<td>19</td>
<td>22</td>
<td>536</td>
<td>552</td>
</tr>
</tbody>
</table>
statistical data that distinguishes between easy and difficult items with the items’ descriptive profiles points to a perhaps more accurate distinction between accessible versus inaccessible items—ones that do or do not do a good job of letting students show what they know and can do in more than one content area.

### Multiple-Measure Activities Yielding Scores Close to Mean Scale Score Locations

Two thirds of the multiple-measure items yielded scale score locations close to the means for both LU and the content area being assessed, suggesting that addressing LU demands did not appreciably impact the difficulty of test items or distort evidence of students’ proficiency in that content area. One typical example of such an
item comes from a task on entrepreneurship, a key concept covered under economics, one of four content outcomes for social studies assessed on MSPAP. The task begins with a preassessment in which students are introduced to the concept of entrepreneurship and brainstorm together realistic jobs for eighth graders. During the course of this task, students complete a number of constructed-response activities in which they identify a business that someone their age might start; they address issues of supply and demand, resources needed, and pricing of the good or service they would offer. As the culminating activity, students write a letter to persuade potential customers to patronize a business they have established:

Now you are going to write to persuade. When you write to persuade, you want to convince someone else to do or think about something the way you do. When you write to persuade, you do the following:

- Decide what your position or stand is on the topic or subject.
- Think of your reasons for that position or stand.
- Think about the possible position or stand of the reader.
- Organize in a logical way the reasons for your position or stand.
- Invite your reader to understand and share your position or stand.

Suppose you find out that other teenagers in your neighborhood have started businesses that offer the same good or service as yours. You want to convince neighbors to use your business rather than your competitors’. Write a letter you can distribute to your neighbors convincing them that your business better meets their needs. Be sure to use what you know about economics and business to convince them that your business is better.

The response space in the student answer book is formatted to include date, salutation, and closing, and consists of fifteen lines; both writing and LU icons appear along the left margin of the response space.

Responses to this activity were scored for social studies, writing, and LU (see Figures 2 through 4). The scale score locations for each area assessed through this activity were close to the means for all items in those areas: social studies, 538 ($M = 533$); writing, 526 ($M = 538$); and LU, 540 ($M = 554$). Despite the placement of this activity at the end of the task, nonattempts were infrequent. Close analysis of item cuing reveals various text features that were typical of items with this sort of statistical profile. The activity includes, as do all those scored for writing, scaffolding to provide

---

1Task security constraints preclude the publication of any portion of tasks not released to the public, including at present all those in the 1996 Maryland School Performance Assessment Program. To protect test security, many activities that appear in this article are analogues created to serve as illustrations and to approximate as closely as possible the essential format and cuing of actual test items.
This response gives evidence of the student’s ability to analyze the effects of supply and demand on the production of goods and services in order to determine what action is required to effect change.

2 = This response gives considerable evidence by including three or more appropriate pieces of information (see answer cues) other than name and phone number.

1 = This response gives adequate evidence by including at least two appropriate pieces of information other than name and phone number.

0 = Other (incorrect or less than adequate evidence)

Answer cue:
Some typical responses may include:

- increased hours of operation
- “new” lower prices
- more extensive selection or services
- high(er) quality of materials
- testimonials
- any other relevant ideas/information based on economics and business operations

FIGURE 2  Social studies activity-specific scoring key. Although in the early years of the Maryland School Performance Assessment Program, some score point descriptors like those in this scoring tool included references to a specific number of elements (reasons, pieces of information, etc.) not explicitly cued for in the corresponding activity, efforts have been taken to eliminate this sort of mismatch. The richness of the information source cues in this particular item appears to have permitted students to meet and exceed the target number of elements identified in the scoring tool, however, as evidenced by the proximity of this item’s scale score location to the content area mean.

guidance on the intended purpose for writing. In this case, students are cued in particular to express their position clearly and to provide reasons to support that opinion. The body of ideas and information that they may draw on for that support is cued for explicitly (“Be sure to use what you know about economics and business”) and embraces test-based information (from previously completed activities) as well as prior knowledge—both content based and personal. Thus, students can tap into an extensive and open-ended range of ideas, anecdotes, and examples to make a compelling case for potential customers to choose their business. In addition, the scenario is credible, and rhetorical specifications are clear and familiar. The activity contains a single topic cue (why your business meets customers’ needs), an action cue for purpose (“write … convincing”), and a support cue (“use”).
Scoring Rule: Language in Use

2 points
☐ Consistently uses word and sentence order and language choices to express meaning with style and tone. Text conveys uniform impression of correctness* and any errors that are present represent risk-taking.

1 point
☐ Sometimes uses word and sentence order and language choices to express meaning with style and tone. Text generally conveys impression of correctness* and errors may or may not represent risk-taking.

0 point
☐ Rarely or never uses word and sentence order and language choices to express meaning with style and tone. Text appears error-ridden.

* correct usage, punctuation, spelling and capitalization

FIGURE 3 Scoring rule: Language in use.

Scoring Rule: Writing to Persuade

2 points
☐ Consistently addresses audience’s needs by identifying a clear position and fully supporting or refuting that position with relevant information. Text is uniformly organized, and language choices often enhance the text.

1 point
☐ Sometimes addresses audience’s needs by identifying a somewhat clear position and partially supporting or refuting that position with relevant information. Text is generally organized, and language choices sometimes enhance the text.

0 point
☐ Rarely or never addresses audience’s needs by failing to identify a clear position or failing to adequately support or refute a position that has been identified. Text lacks organization, and language choices seldom, if ever, enhance the text.

FIGURE 4 Scoring rule: Writing to persuade.
Review of the three different scoring tools used to evaluate responses to this activity make evident that skills and knowledge contributing to proficiency in each area have not been confounded. Students may write compelling persuasive letters that receive a “2” for writing but are devoid of evidence of familiarity with key economics concepts and, thus, receive a “0” for social studies. Alternatively, responses may both exhibit correct sentence formation and mechanics and include numerous economically based ideas (thus earning a “2” for both LU and social studies), but not articulate a position and support that position with reasons (and thus receive a “0” for writing). A third of many permutations (illustrated in Figure 5) is a response that receives full credit for social studies and writing, whereas only partial credit for LU.

Striking features of this item are its purposeful repetition in directions and its clarity. The concepts of “needs and wants” and “goods and services” that it addresses have been the focus of earlier activities within the task and so are familiar to students. The rhetorical situation is an authentic one, both in terms of context and relation to audience. The activity is open-ended, giving students a wide range of choices rather than demanding they present a single “right” answer, and enabling them to draw on both academic and personal information sources to develop their ideas. Cues do not appear to function at cross purposes with, or to obscure others. In fact, rather than creating tension between the demands of the content area (social studies) and of writing or LU, this activity is an example of one in which fulfilling the content area expectations supports the writing, just as the writing provides an authentic opportunity to employ the content area concepts and knowledge. We might say that this item, like many of the multiple-measure items on MSPAP across content areas, is accessible—that is, the cuing makes clear to students what they need to do, providing an unimpeded opportunity for students to demonstrate proficiency in the multiple areas the item is intended to assess.

Typically, across all content areas, items for which the scale score location was relatively close to the mean for all items in that area exhibited similar cuing patterns. These activities tended to be characterized by a clear presentation of content and rhetorical demands. In addition, when measuring content and LU, but not writing, directions were very streamlined. Audience and purpose were simply named without extensive contextualization within the item itself, enabling students to focus appropriately on content elements within a clear and correct written response. These activities tended to limit the number of cuing verbs and identified a focused topic. Such items were in evidence across content areas, even in mathematics content and science, as illustrated by the following Grade 3 science item:

Think about the predictions [about physical properties] you made in the chart on page X for the ping pong ball and the golf ball. Use what you know about the ping pong ball and the golf ball to write a note to a classmate to explain why you made your predictions.
Activities like this one did not require a rereading to decipher what was expected. Information source cues were often explicit and, if test based, also permitted prior knowledge, which often took the form of personal ideas, interpretations, or opinions (e.g., “What would you predict?” or “Would you recommend?”).

Multiple-Measure Items With Outlying Scale Score Locations

With two exceptions, items flagged as outliers had higher than mean scale score locations, suggesting that they were more difficult than other items for that area. Variation in scale score location, suggesting a range in item difficulty, is a characteristic of well-designed tests and is something to be expected and indeed welcomed. Rather than assuming that outliers were problematic, therefore, we sought ways of differentiating between items that were truly more or less difficult due to the content and concepts being assessed and those that might distort an estimation of students’ abilities because of particular text features of some multiple-measure items.

*Items with outlying content area scale score locations.* For some items, the scale score location for the content area was higher than the mean, whereas it
was not higher for writing or LU. Such items did not appear to create a ceiling for writing or LU scores even when content knowledge was limited, as in the following example from Grade 5 scored for reading and LU:

Write a note to your teacher explaining which reading [text selection] you think would be more helpful to other fifth-grade classes who are designing a decorative geometric border around the gym. Use information from the reading you chose to support your choice.

Prior to this activity, students had been introduced to the context (using geometry to create a border design) and had read two selections: one defining various geometric shapes and another describing these shapes as they appear in the natural world. Although the LU item location (537) was close to the mean (535), the location for reading (565) was much higher ($M = 520$). We surmise that this difference legitimately reflects that evidence of proficiency in reading required text-based information (“Use information from the reading”), whereas students might draw on any prior knowledge, both content and personal (“which you think”) to construct a response. Even if they could marshal little or no evidence of construction of meaning from the reading selections, students had ample opportunity to develop sufficient text to demonstrate proficiency in LU.

Items that elicit this combination of typical writing and/or LU scale score locations and significantly higher scale score locations for the content area being assessed (accurately reflecting students’ relatively weaker content knowledge and skill) are generally characterized by various cuing and format features. The information source cues allow for support with prior personal knowledge that takes the form of an opinion or impression as well as with test-based information that involved personal observation or hands-on engagement. Students were enabled to find something to write about even if they were unable to address particular content area demands accurately or appropriately for the particular discipline being assessed. (See Figure 6 for an illustration of student responses to an item that functions in this manner.) This sort of item may be considered effective and accessible because content difficulty does not confound performance in other areas.

Sometimes, however, typical LU scale score locations coupled with higher than usual content scale score locations characterize items that generate text successfully but that may, in the process, confound evidence of content knowledge. In one Grade 3 mathematics task, for example, students completed a number of activities dealing with patterns and measurement as part of a scenario in which a company is holding a contest for a container design. Midway through the task, they are cued as follows:

Write a note to the company’s artistic director to send along with your container design. Explain to the director which shapes you used in your design
and how they are used to make your pattern. Be sure to include the names of the shapes and how you used them to make the pattern on your container.

Responses to this activity were scored for both mathematics content and process, along with LU. Although the LU scale score location (543) was close to the mean (542), the math content scale score location (589, $M = 536$) and math process scale score (570, $M = 544$) were not. Examination of this item reveals a number of problematic features. To begin with, the activity does not make clear to students what writing this letter has to do with entering the contest; thus, the motivation and context for the item is not clear. The rhetorical situation is also likely to be confusing to students, for it is hard to imagine a credible situation in which someone would “name shapes” or explain their use in a letter to a company’s director. Unlike items in which LU or writing cuing does not interfere with content cues, and
like other problematic ones, this activity is cluttered and contradictory. There is a conflict in cuing verbs because the activity calls for an explanation but really simply requires naming and describing. The information source (“which shapes you used”) also miscues because, although it directs students to consider a previous, specific, test-based activity, it only implies rather than makes explicit that what is being assessed is knowledge of mathematical patterns and geometric shapes. In other words, the cuing neither elicits the language of mathematics that would give evidence of proficiency in the content outcomes being assessed nor does it make clear what skills and processes are being measured. Far more effective might have been cuing to “Use what you know about patterns in mathematics to describe the pattern you created on your container.” We conclude that this is not a difficult item but, instead, a problematic one in which students do not have a good opportunity to show what they know and can do. The fact that other items from this test form addressing the same mathematics outcomes and indicators actually yielded scale score locations below the means supports this conclusion.

Similar patterns in relative difficulty suggest additional features that may confound content area scores, as in the case of the following illustrative item appearing at the end of a Grade 8 task that required students to do a number of activities related to a chart and to call on extensive prior knowledge:

Science Week is celebrated each spring to remind students of the impact science has on our everyday lives. One of the activities being planned for this year’s Science Week is a school newsletter to inform students of scientific principles that are around us every day. When you write to inform, you share with others what you know about a subject. When you write to inform, you do the following:

• Think about what the reader already knows about a subject.
• Think about what the reader needs to learn about a subject.
• Put your information in a logical order.
• Use examples, definitions, and descriptions to make the information more clear to your reader.

You have been asked by your teacher to write an article on motion for the Science Week newsletter. Before you write, think about:

• The elements at work when something is moved by something else.
• The relationship between mass and acceleration.
• How this relationship applies to a situation in people’s daily lives (such as the force to get a loaded grocery cart moving).

Write an article for your school’s Science Week newsletter that explains how Newton’s second law of motion applies to a situation in people’s daily lives.
As in the container design task, LU (560) and writing (550) scores are close to the means for those areas (\(M = 544\) and \(M = 538\), respectively), whereas the item scale score location for science (586, \(M = 544\)) is not. The presumption that this is simply a more difficult science item is again called into question by comparison of item scale score locations for other activities in this task measuring the same outcomes (processes of science and application of science). Although it is certainly conceivable that there could be varied degrees of difficulty among items measuring the same outcome, a close look at item format and language delivers a more plausible explanation of the outlying scale score location.

This particular item is so laden with cross cues that it is difficult to read it and step away with a clear notion of what to do. Whereas many multiple-measure items create the conditions whereby students are induced to write about something, even if their content knowledge is weak, this item seemed to shut down responses altogether. Nonresponses were high—20% of all students taking this test form (perhaps a function as well of this being the last activity in the last task on the last day), as were true “0” scores (ranging from 37% for LU, to 45.2% for writing, and to 47.4% for science). This activity is characteristic of those in which the demands of one area are in tension with the demands of the others. Unlike the entrepreneurship activity discussed earlier, in which the context is established in a believable and timely way, in this case the fiction of a newsletter for Science Week is “tacked on” and has not been prepared for by the previous activities within the task (an unusual departure from current test specifications that call for all writing activities to be clearly introduced at the outset of the task). Especially troublesome is the language at the beginning of the activity that prepares students to think about science in general terms rather than the specific concepts being assessed. It is only in the very last line of the item (after students have read over 150 words) that they are directed to explain Newton’s second law of motion. It is hardly surprising that some students focused on the rhetorical expectation that they make the topic of motion interesting to readers, failing to address Newton’s law in detail—if at all—whereas others simply reiterated Newton’s law without fully explaining its application in daily life.

Both this and a number of other activities suggested the caveat that sometimes “two’s company, three’s a crowd”—that is, that it was easier to create activities that effectively measured proficiency in a single content area and LU than in a content area, LU, and also writing. Typically, in problematic cases, the students were not returned via content cues to the essential substance required in their responses. Often, students seemed to become so caught up in the rhetorical demands established by the writing scaffold and reiteration of specifications that the content focus was obscured. A case in point is the following activity from a Grade 3 task:

Now you will be writing to inform others about the airplane you have designed. When you write to inform, you share with others what you have learned about a topic. When you write to inform, you should remember to:
1. **Think about the facts the reader needs to know.**
2. **Put your facts in the best order.**
3. **Use examples and descriptions to make the information clear to your reader.**

Now that you have made the decisions about the material and shape of your airplane, you are ready to enter it in the airplane show. Visitors to the show need some information about your airplane. Prepare information for a display card which the visitors can read. Name your airplane. Tell why you designed your airplane the way you did and why you think it will fly. Use information from what you know and what you have learned from this task to support your ideas. [This text is followed in the student’s answer book by a graphic consisting of a lined and bordered box labeled “Display Card.” The first line contains the words “Name of Airplane”].

Item scale scores for LU (535, $M = 542$) and writing (542, $M = 560$) were again close to the mean (and if anything, suggesting a somewhat easier than usual item), whereas for science, the item scale score (564) was much higher than the mean (512). This item is one of a number of instances in which students may have been “killed with kindness”; that is, in an effort to make an activity appealing and its varied demands more clear, the scaffolding may have gotten in the way or background cues may have clouded what was really being measured in the response. Although students are cued to “use information from what you know and what you have learned,” it is never indicated what kind of information or specific knowledge is required (e.g., what you know about aerodynamics or what you have learned about the relation between materials, form, and motion). This lack of specificity signaling what content or concept is being measured is especially striking given the directness with which students are cued to include irrelevant information such as the name of their airplane.

This activity, and others resembling it, makes clear that what makes an item challenging may be attributed as much to its text features as to content area. We must revisit any initial impression that some content areas—social studies and reading—by their very nature lend themselves to multiple measures more so than others—science and mathematics content. Rather, what we discovered is that outlying science and mathematics items were frequently characterized by vague information source cues, whereas reading and social studies items typically made clear precisely what text elements (use words and phrases from the poem) or concepts and content knowledge (use what you know about the availability of resources and the production of goods) students needed to tap into to develop their response.

Unlike the airplane design item and others in which students were killed with kindness, some activities were entirely too compressed, giving students no opportunity to reflect and make the necessary connections between things they had done.
and learned and the demands of the activity at hand. A simple process of “parsing out” such activities might enhance their accessibility and permit them to function effectively as sequentially scored items. Therefore, for example, we could take one activity (Grade 5) and revise it as follows:

[Original] In one or more paragraphs, write instructions that classmates could use to check the accuracy of the repaired barometer. Use what you know and what you have learned from this barometer activity to create your instructions.

[Revised] You have repaired the barometer and believe it will work accurately now. How could your classmates check the accuracy of the barometer? Use what you know and what you have learned from this task to explain to your classmates what they could do to check the accuracy of the barometer.

It is worth noting that this item was revised after the 1996 administration both by parsing out instructions and cuing more clearly for what was being assessed (an understanding of how to use a barometer to measure air pressure). When the revised version was used in the following year’s test, the item scale score location for science changed from 616 to 513. Although one hopes that improved instruction accounts at least somewhat for this difference, the increased accessibility of the item must also have contributed to students’ better performance.

Items with outlying LU scale score locations. Although the majority of items designated as outliers were flagged because of discrepant content area item scale score locations, a small number of items (six) either were outliers in all areas measured or in LU alone. Again, this pattern sometimes appears to reflect the true difficulty of content concepts and knowledge, whereas in other instances, text features of the item related to content cuing appear to have distorted evidence of proficiency in LU. Consider, for example, the following Grade 3 activity in which, after estimating the number of triangles needed to complete the border around a bulletin board represented in a graphic, students were asked the following:

Look at your estimate. Write an explanation for a classmate to tell how you estimated the number of triangles needed to finish the border.

For this item, the math process scale score location of 535 was close to the mean (539), unlike the scale score locations for math content (550, $M = 526$) and LU (563, $M = 542$). Rather than being fundamentally difficult, this item appears to be one in which the combination of specific information source cues (calling for test-based and prior content knowledge) and a restricted opportunity for development very likely limited what students were prompted to do and caused a false ceiling for LU. Activities of this sort typically had a single or very limited range of acceptable re-
responses for the content area being measured, and only permitted students the narrowest of choices of how to present an explanation. Beyond naming or identifying something, there was little more students needed to, or could, say—in effect, they were “short circuited.” Typically, content measures in these instances were derived from dichotomous (0 to 1) scales, which should have perhaps tipped off task writers that the scope of the response demands would not facilitate a measure of writing or LU scored on a 0 to 2 scale. Furthermore, these items often demonstrated a mismatch in which the rhetorical demands (e.g., to make a recommendation in a speech to a committee) did not fit the content demands (articulating the mathematical procedures and calculations used to arrive at that position), or in which audience is simply tacked on without creating a credible context or motivation. These items provide additional evidence that it is not content area in and of itself, but rather the construction of certain items, particularly those assessing math content, that leads to the greater prevalence of outliers among multiple-measure items in those areas. Often, these outliers are not truly difficult, but rather are problematic in the effects of their cuing on the development of students’ responses.

CONCLUSIONS

This study supports a broadened approach to performance assessment item analysis as a critical component of performance task and test development. As with the creation of effective writing prompts (Ruth & Murphy, 1988), a full understanding of the behavior of constructed-response items across content areas requires that we go beyond item statistics to the finer grained descriptive analysis of the actual language and format of such activities. This process, in turn, causes us to reexamine the terms by which we categorize assessment items. Using statistical data to distinguish between easy versus difficult items may be obfuscating a less obvious but equally important distinction between accessible versus inaccessible items—items that do or do not do a good job of letting students show what they know and can do. Responsible test development and use require that we try to distinguish between items that are simply harder and thus successfully addressed only by the more able student from those that may merely be poorly worded and formatted and that only some students (not necessarily the more able ones) have been able to “unlock.” If we reorient our thinking to focus on accessibility rather than difficulty, we discover that the features that characterize problematic multiple-measure items and distinguish them from accessible ones can be both anticipated and avoided (see Figure 7). Certainly, many of these features and the guidelines they imply mirror already familiar principles for the development of effective performance tasks. Although the literature of performance assessment reiterates that tasks should be authentic and clear, for example, we believe that it is only on the micro level of cuing that we can un-
understand how to achieve these widely shared goals. The framework for analysis we offer, addressing such concerns as the identification of information sources and quantity of cuing verbs, provides a systematic representation of the work students must do to “unpack” any item. With any activity, students ask themselves, “What do I need to do?” and “What knowledge and skills must I demonstrate?” Perhaps least often addressed in the assessment literature, but critical to students’ success, is a third question: “What sources am I to access in order to do this?” By adopting the perspective of respondents, and analyzing the extent to which items give clear and noncontradictory answers to these questions, we refine the principles for good task and test design and extend them both to multiple-measure items and to other forms of constructed-response items as well.

<table>
<thead>
<tr>
<th>A More Accessible Item...</th>
<th>A Less Accessible Item...</th>
</tr>
</thead>
<tbody>
<tr>
<td>makes explicit the specific content area skills and/or concepts being assessed</td>
<td>fails to include or is ambiguous about the content area skills and/or concepts being assessed</td>
</tr>
<tr>
<td>makes clear the range, focus, and/or quantity of elements (ideas, information, details) that are necessary for a successful response</td>
<td>fails to make clear the elements that are necessary for a successful response and/or obscures this information with extraneous detail</td>
</tr>
<tr>
<td>is uncluttered, clear, direct, and focused, with good recall capability</td>
<td>is crowded with too many content topics, cuing verbs, or syntactic complexity</td>
</tr>
<tr>
<td>permits open-ended development</td>
<td>requires single or limited “right answer”</td>
</tr>
<tr>
<td>makes information source explicit (“Use what you’ve learned about ___ from ___”)</td>
<td>only implies or does not identify information source (e.g., “Use what you know”)</td>
</tr>
<tr>
<td>allows student to draw upon personal, prior knowledge and/or multiple information sources at hand (e.g., written text, drawings and photographs, diagrams, graphs)</td>
<td>restricts response to highly content-specific and/or a narrow range of test-based information sources (e.g., knowledge of a particular formula or a single graph)</td>
</tr>
<tr>
<td>establishes a rhetorical situation that is engaging and believable</td>
<td>establishes a rhetorical situation that is contrived</td>
</tr>
<tr>
<td>creates a good fit between rhetorical and content demands, one in which the rhetorical situation supports development of content-based ideas/information</td>
<td>creates a mis-fit between rhetorical and content demands</td>
</tr>
<tr>
<td>creates or calls up an audience-writer relationship that is authentic, plausible, and developmentally appropriate</td>
<td>creates inauthentic audience-writer relationship or one which is socially/culturally implausible or developmentally inappropriate</td>
</tr>
</tbody>
</table>

FIGURE 7 Characteristics of more and less accessible multiple-measure items.
Performance assessments that rely on paper-and-pencil, constructed-response items require that we recognize that development is an art, albeit one suffused by the science of statistics. In the execution of their art, test developers have at their disposal, and ought to use, the wide range of resources available to them—not only statistical data, but careful readings of student responses and, perhaps most important, the close and critical analysis of the effects of item language and format on student performance. This is already a challenge when dealing with a single domain or content area, and even more so when attempting to obtain measures of different areas from the same assessment activity. In pursuing this goal, however, test developers enter a new frontier in fulfilling the promise of integrated performance assessment.

ACKNOWLEDGMENTS

An earlier version of this article was presented at the annual meeting of the American Educational Research Association, March 25, 1997. We thank Hillary Michaels for assistance on an early draft of this article as well as the anonymous reviewers who provided valuable suggestions for the revision.

REFERENCES


