Psychometric Considerations for the Next Generation of Performance Assessment

Authors:
Tim Davey, Educational Testing Service
Steve Ferrara, Pearson
Paul W. Holland, Emeritus, University of California, Berkeley
Rich Shavelson (Chair), Emeritus, Stanford University
Noreen M. Webb, University of California, Los Angeles
Lauress L. Wise, Human Resources Research Organization

Commissioned by:
Center for K–12 Assessment & Performance Management at ETS
Educational Testing Service

Supported by:
Charlene G. Tucker, Project Manager

Acknowledgments

The study group members want to acknowledge the contributions of many individuals whose thinking, contributions, and support made this work possible.

Center for K-12 Assessment & Performance Management at ETS: For believing in this project and securing all the resources and support that made it possible, we thank many individuals, but especially:

- Pascal D. Forgione, Jr., Distinguished Presidential Scholar and Executive Director
- Nancy Doorey, Director of Programs

Presenters: To provide currency on particular related topics, we invited presentations from top scholars in the field. We thank the following individuals for preparing presentations and sharing their expertise:

- Brian Clauser, National Board of Medical Examiners
- Peter Foltz, Pearson
- Aurora Graf, Educational Testing Service
- Yigal Rosen, Pearson
- Matthias von Davier, Educational Testing Service

State stakeholders: For providing guidance as our work was launched, for reviewing our drafts, and for generally helping us make sure our work is relevant to the field, we thank:

- Jeffrey Hauger, New Jersey Department of Education
- Pete Goldschidt, New Mexico Public Education Department
- Juan D'Brot, West Virginia Department of Education
- Garron Gianopulos, North Carolina Department of Public Instruction

Consortia stakeholders: For helping us understand the technical challenges facing the Race to the Top assessment consortia as they incorporate performance assessment into mainstream K-12 assessment, for their guidance as we were launching our project, and for their reviews of our drafts, we thank:

- Enis Dogan, Partnership for the Assessment of Readiness for College and Careers
- Marty McCall, Smarter Balanced Assessment Consortium

Technical reviewers: For providing thoughtful and thorough reviews of our draft which undoubtedly contributed to its accuracy and clarity, we thank:

- Suzanne Lane, University of Pittsburgh
- Matthias von Davier, Educational Testing Service

Additional support: We additionally thank the following people for all the things we didn't have to worry about:

- Charlene Tucker, for organizing our work and making sure we had the resources we needed
- Rosemary Calvert, for graciously hosting our meeting at the beautiful San Francisco ETS office

It was our pleasure to work with and learn with you all.

Chapters

I. Overview of Psychometric Considerations for Next Generation Performance Assessment 5
II. Definition of Performance Assessment 17
III. Performance Assessment: Comparability of Individual-Level Scores 38
IV. Performance Assessment: Comparability of Groupwork Scores..... 57
V. Modeling, Dimensionality, and Weighting of Performance-Task Scores 82
VI. Challenges and Recommendations..... 97

I. Overview of Psychometric Considerations for Next Generation Performance Assessment[1]

Introduction 6
Argument for Performance Assessment … in Brief 6
The Elephant in the Room 6
Purpose 7
Scope 7
Summative, Formative, and Embedded Assessment 8
Assessment Versus Teaching-Learning 9
What This Paper Does Not Cover 9
Audience 10
What Is Performance Assessment? 10
Big Issues Confronting Performance Assessment 11
Overview of the Report 13
Chapter II Definition of Performance Assessment 13
Chapter III. Performance Assessment: Comparability of Individual-Level Scores 13
Chapter IV. Performance Assessment: Comparability of Groupwork Scores 14
Chapter V. Modeling, Dimensionality, and Weighting of Performance-Task Scores 14
Chapter VI. Challenges and Recommendations 15
References 15

[1] Rich Shavelson was lead author for Chapter I.

Introduction

Gorin and Mislevy (2013) presented a paper—*Inherent Measurement Challenges in the Next Generation Science Standards for Both Formative and Summative Assessment*—at the Research Symposium on Science Assessment. They unknowingly set the agenda for the study reported here. According to Gorin and Mislevy:

> For the summative accountability use case in which reliability and generalizability are of most importance, policy makers and educators need to understand what the limits of our current psychometric capabilities are. Simultaneously, psychometricians would be well advised to broaden their perspectives to embrace alternative approaches to statistical modeling and scoring that may not be as familiar or comfortable to us, but are more likely to provide key stakeholders with the type of information that they so desperately need to improve science education. (2013, p. 24)

Argument for Performance Assessment … in Brief

Our work began on a familiar premise: Performance assessment provides valuable information that traditional assessment types cannot—observable performance using higher order critical thinking with generic and domain-specific content in real-world contexts (e.g., Darling-Hammond & Adamson, 2010; Frederiksen, 1984; Frederiksen & Collins, 1989; Johnson, Penny, & Gordon, 2009; Wigdor & Green, 1991). Both the Common Core State Standards (CCSS) and the Next Generation Science Standards (NGSS) have made the case that performance assessment adds value in both the types of complex skills that can be measured and the types of educational strategies it reinforces and informs. Justification for the use of performance assessment, then, rests on a number of grounds such as performance assessment (a) signals the importance of both higher order thinking and using that thinking *to do* in a practical real-world context, (b) measures such thinking and doing in ways traditional selected-response tests cannot, and consequently (c) complements traditional assessment to measure more fully the competence or construct of interest. Moreover, performance assessment may also lead to increased learning (e.g., Lane, Parke, & Stone, 2002; Stone & Lane, 2003).

Indeed, there are occasions when common sense dictates the use of performance assessment in addition to traditional assessment (e.g., Wigdor & Green, 1991). No justification is needed for the performance assessment included in driver's examinations. Answering multiple-choice questions regarding driving is not the same as demonstrating driving competence in a car. In a similar vein, answering multiple-choice questions about carrying out a well-controlled science experiment is not the same as actually conducting that experiment (e.g., Baxter & Shavelson, 1994).

The Elephant in the Room

For all the potential benefits of performance assessment, such assessment presents substantial conceptual and measurement challenges, whether used in education, certification, civilian work, or the military. As Gorin and Mislevy (2013) pointed out, performance assessment presents significant psychometric challenges—in *task design and scoring*, *psychometric modeling*, and *practical and logistical feasibility*—that need to be addressed to ensure the defensibility of incorporating this type of assessment into a system of accountability.

These challenges are like the proverbial elephant in the room—the elephant is there but no one wants to acknowledge or talk about the elephant. The conceptual

challenges include defining just what performance assessment is. What might be included as performance assessment and what falls outside the bounds of the definition? What knowledge and skills do performance assessments tap that other formats cannot?

The measurement challenges are even more daunting. Typically performance assessments measure individual or group extended performance (in contrast to 30 second multiple-choice items), producing relatively few responses for scoring, on items that defy the usual assumptions of local independence (items are linked by a common problem or task) and unidimensionality (higher order multidimensional abilities and skills involved), providing extended responses that need to be scored, at least initially, by human raters (and perhaps later by machine) and may be embedded in classroom or work activities such that authorship of the performance or test security may be questionable (e.g., teacher or parent supports, training on job-performance tasks).

Purpose

The purpose of this report is to shine a light on the elephant in the room, exposing the conceptual and methodological challenges of performance assessment. We do this by reviewing what is known about performance assessment and recent psychometric developments that might address some of the challenges, and by identifying areas for new developments. To this end, the next chapter, Chapter II, takes up the definition of performance assessment. The two chapters that follow Chapter II enumerate challenges and possible remedies in producing reliable, comparable scores on performance assessments. One chapter focuses on an individual's performance (Chapter III) and the other on performance in and of a group (Chapter IV). Chapter V takes responses from performance assessments and discusses alternative ways of modeling them to produce reliable and valid scores. And Chapter VI, the final chapter, summarizes the issues identified in the report and recommendations for addressing them along with speculations for future psychometric developments.

Scope

This report focuses on performance assessment in large-scale testing programs, primarily in education. To this end, most examples and literature are drawn from education. But the report has a more ambitious goal: speaking to other fields such as certification, performance measurement in civilian and military jobs, or in sports where human performance generally is of concern. Such testing programs, then, might assess achievement and holding schools accountable as Race to the Top intends. Such performance-assessment programs might assess college readiness in college admissions testing such as that carried out by the Council for Aid to Education with its Collegiate Learning Assessment (e.g., Shavelson, 2010). Performance assessment might be used to assess competence for large-scale certification or licensing as is done with the California Bar Examination.[2] Or performance assessment might be used in job-performance testing such as that carried out in industry (e.g., assessment centers—e.g., Collins & Hartog, 2011). Consequently, examples and scientific evidence regarding performance assessment are drawn more broadly than from education. Indeed, given the gaps in education research on performance assessment for individuals and groups, we found it essential to look at what is known in other fields.

[2] See http://www.calbar.ca.gov/

One concern, then, might be that this report is not tailored specifically to the assessment agenda for the new education standards (i.e., CCSS or NGSS). However, the charge to the group writing this report, one we agreed with, was to go beyond the current education assessment environment and speak more broadly. The report is not intended to and cannot replace the work of the technical advisory committees attached to the effort to assess the new education standards. We believe that such committees will find the breadth of this report useful in gaining insights into how to meet the challenges of their educational assessments.

Summative, Formative, and Embedded Assessment

In educational testing, a distinction is made between summative and formative assessment. Because those involved in education policy, practice, and assessment form a major audience for this report, it seemed appropriate to distinguish among these two forms of educational assessment. Summative assessment covers a broad range of uses such as school accountability or high school graduation, college admissions, long-term instructional improvement, certification and licensure, and job performance. Summative assessment, then, is aimed at taking the temperature of, for example, the education system to inform policy and curriculum. It is typically carried out on a large-scale, with stakes attached, and provides a broad assessment of what students know and can do at the end of a given activity or period of time (e.g., eighth grade, medical education)—a sort of summary judgment. Examples of such educational assessments include mathematics and English language arts achievement assessments across the states in response to No Child Left Behind legislation, the National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA), and the soon to be launched Race to the Top assessments. Examples outside education include professional certification and licensure tests (e.g., medicine, law, driving). This report focuses on the psychometric challenges of using performance assessment in the context of summative assessment.

Summative assessment can also be used for instructional improvement, even though the temperature taken is at a great distance from the classroom. Summative assessment can show areas of curriculum and teaching that are strong and weak, and it can stimulate conversations and pedagogical actions for the future. However, such information is not going to inform teachers' everyday decisions about how best to teach their particular class of students the particular content at a particular moment in time. Relevant information is needed on the fly from formative assessment.

Perhaps noticeably absent in this report, then, is the use of performance assessment to inform instructional decisions—*formative assessment*. Formative assessment is intended to index the gap between where students are and where they are expected to be with respect to some learning outcome (e.g., problem solve with linear equations). The intent is to inform the student of goals and progress toward them on the one hand and to inform teachers of the gap so they can devise instruction to close the gap with constructive feedback to the student as to how to improve (e.g., Ruiz-Primo, Furtak, Ayala, Yin, & Shavelson, 2010).

We have avoided formative assessment not because it is unimportant but because such assessment warrants a separate report on its own. Some performance assessment challenges are not as great with formative assessment as in large-scale testing (e.g., high reliability is not a necessity as the teacher can adapt to the information provided by the performance assessment if not reliable); other challenges remain such as task construction, scoring time, and logistics.

Of course, there is an exception to every rule, and in our report, *embedded assessment* is the exception. In an attempt to (a) provide sufficient time for extended performance assessment and (b) provide feedback to classrooms almost immediately, some performance assessment can be embedded in a classroom (or work site). So embedded assessment is like formative assessment in that performance assessment is embedded within a classroom. Embedded assessment is also like formative assessment in that the information provided can inform both student and teacher of gaps. However, embedded assessment is unlike formative assessment in that the tasks are set externally as part of a summative assessment system; the psychometric requirements for embedded summative assessment are roughly the same as those for the on-demand version in a testing center or classroom.

Of particular concern with embedded assessment is that such assessments are not given under well-controlled (standardized) administration conditions. Authorship is at issue. Given the demands of high-stakes testing, the question of who was the author of the test responses is critical in drawing inferences about individual or group performance. In particular, concern attaches to the kinds of support that teachers and others might provide students, raising the question of whose performance is being evaluated. In some cases, such as portfolios or projects, a question of authenticity may arise as authorship may be particularly difficult to ascertain.

We believe that what we have to say about large-scale, summative assessment also applies to what is called embedded assessment. However, we will not speak directly to embedded assessment in the remainder of this report.

Assessment Versus Teaching-Learning

A good assessment task is a good teaching task and vice versa. It is sometimes difficult to distinguish the two. It is also easy to slip into a discussion of instruction and the corresponding literature instead of assessment. This is especially the case with performance assessment. Our focus is on assessment, not teaching-learning (although, of course, all three are intimately intertwined). Assessment, especially summative assessment, comes with certain requirements that distinguish it from instruction. Most notably, assessment has to stand alone without support of a teacher, trainer, or test administrator. A teacher or test administrator cannot step in and fix an ambiguous instruction or task as a teacher might do during formative assessment. The stand-alone requirement makes the development of performance assessment for summative purposes quite challenging especially because they are somewhat lengthy. Consequently we spend some time talking about performance assessment plans or blueprints. We do not, however, spend time on instructional uses of performance assessment information.

What This Paper Does Not Cover

Perhaps it is somewhat surprising, including surprising to us, that we do not have a chapter devoted to information technology. Such a chapter might be expected because, among other things, information technology has the capacity to support (a) task development/delivery, (b) capture and analysis of process data, and (c) scoring techniques. The role of technology in performance assessment is a very important topic for a future report.

This report also does not cover psychometric challenges of performance-assessment accommodations/accessibility for English language learners and students with disabilities. That topic also deserves a report of its own.

Audience

This report aims to speak directly to test-development and measurement experts found in state departments of education and labor, in professional associations, and in testing companies. To this end, we assume a certain degree of reader sophistication with terminology, concepts, and statistical/psychometric modeling, but attempt to strike a balance avoiding technical matters including equations. While it is not mathematically rich, this report draws largely on and at times extends well-established psychometric concepts and methods as well as introduces some relatively new ones. To the most sophisticated psychometricians, the presentation will not be technical enough, although we hope there are a few kernels of wisdom that they will find useful. To others involved in testing, the level of sophistication here might be a stretch in places, but we hope that we have struck a balance so that they get the gist of the argument put forth. Finally, to managers and policy makers, we suspect what is presented here will be too much in length and technical jargon.

We have written a short companion piece that presents the ideas and technical issues in an accessible way to a broad audience of policy makers, managers, educators, and the interested public. The companion document, *Psychometric Considerations for Performance Assessment With Implications for Policy and Practice,* provides a synoptic translation of the present document for a wide audience interested in the use of performance assessment on a large scale. In the companion document, key concepts and procedures are defined and important points and findings are translated into language and ideas accessible to the wider audience of policymakers and practitioners.

In the remainder of this chapter, we highlight the topics covered in the subsequent chapters. For readers wishing a quick read, Chapters I and VI provide the gist of this report.

What Is Performance Assessment?

A performance assessment (sometimes called a work sample when assessing job performance), as defined in this report, is an activity or set of activities that requires test takers, either individually or in groups, to generate products or performances in response to a complex, most often real-world task. These products and performances provide observable evidence bearing on test takers' knowledge, skills, and abilities—their competencies—in completing the assessment (e.g., Shavelson, 2013). Such assessments as science performance assessments, essays using informative documents, portfolios, computer simulations, projects, and demonstrations may be considered forms of performance assessment.

More generally, and the *only equation* in the report, we might think of a performance assessment (PA) as composed of performance tasks (PT) and short-answer items (SA) and multiple-choice items (MC):

$$PA = PT_1^{+} + PT_2 \ldots + PT_k + SA_1^{*} + SA_2^{*} + \ldots + SA_m^{*} + MC_1^{*} + MC_2^{*} + \ldots + MC_p^{*}.$$

A performance assessment, then, is composed of at least one performance task (science investigation, essay, portfolio item). The plus (+) in PT_1^{+} indicates that at least one performance task is mandatory to have a performance assessment. The remaining terms in the equation are optional. The asterisk (*) associated with SA and MC indicates that such items are permitted only if they are directly *connected to* the overarching performance assessment rather than to general content (or other) knowledge. That is, such items might provide information as to whether a person

understood a written document that was part of the assessment or could follow a procedure needed for the assessment. We suspect that most performance assessments will be composed of several performance tasks and selected-response and multiple-choice items.

A science performance assessment, for example, might present examinees, individually or in groups, with a complex task such as determining the contents of an electric mystery box that might include one or more batteries, wire, bulb, some combination. or nothing (Figure 1-1). Note that there are multiple mystery boxes (performance tasks) and test takers answer a short-response question for each box—how did you know what was inside the box? Test takers hook up external circuits with wires, batteries, and bulbs to determine the content of a mystery box. The task is complex. It is something (junior) scientists might do, and it requires facts, concepts, mental models, and manual skills to carry out. The response involves both a description of the contents of the box and a justification for the assertion as to what is inside. Student performance is scored for both accuracy and justification (e.g., Ruiz-Primo, Baxter, & Shavelson, 1993).

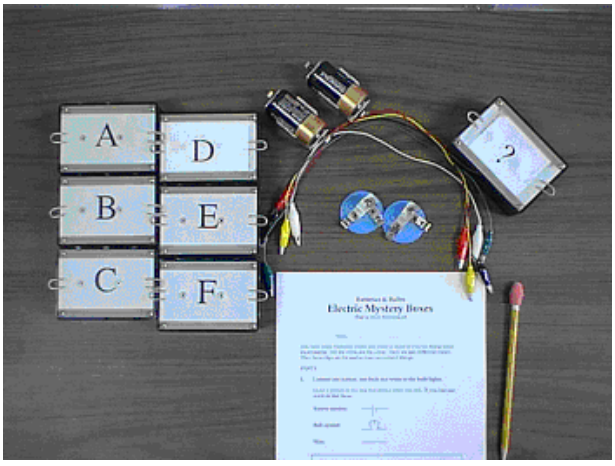

Figure 1-1. Electric mysteries performance assessment.
(See http://web.stanford.edu/dept/SUSE/SEAL/ for more information.)

More generally, a performance assessment requires test takers to produce products and/or performances in response to a complex, typically real-world task. The task typically emulates a context beyond the testing situation—a criterion situation for which we wish to know something about the test takers' performance competence. Completing the task requires the application of higher order (complex) knowledge, skills, and reasoning. The test taker's performance is scored dichotomously 0,1 reflecting, for example, right/wrong, or ordinally (e.g., 1-6) reflecting accuracy, completeness, argumentation and justification, effectiveness, or with some other scale such as time to perform, errors in performing, percentage of key terms used, and the like.

Big Issues Confronting Performance Assessment

There are several, overriding issues confronting the use of performance assessment that devolve into many smaller issues. First, and most problematic, is that performance assessment typically involves complex, real-world tasks and responses that take considerably more time to carry out than do traditional selected-response (e.g., multiple-choice) items. As a result of the time requirement, these performance

assessments typically produce a far smaller *independent* sample of behavior than traditional tests. Moreover some test takers perform well on one task while other test takers perform well on a different task. This examinee-by-task interaction produces large task sampling variability. To obtain reliable scores, a large number of tasks is needed and the cost in dollars, time, (standardized) administration, logistics, and the like is substantial; fewer tasks lead to lower reliabilities (e.g., Gao, Shavelson, & Baxter, 1994; Shavelson, Baxter, & Gao, 1993).

A second issue lies in scoring performance assessment. First, a decision needs to be made as to what constitutes a scorable unit. Is *time* or *error rate* a scorable unit? Or is accuracy the unit? Or is solution-path or process the scorable unit? The definition of what is to be measured provides the criteria for such a decision. For example, if the assessment is intended to measure mathematical fluency, *time and errors* might be the relevant measures. But time probably would not be relevant if the intent is to measure *complex problem solving*.

Second, performance assessment typically produces extended constructed responses (e.g., written, oral) that are evaluated, at least initially, by human raters. These assessments produce myriad score types (e.g., 0, 1; 0–6, time, errors, percentage), many of which are ordinal and their underlying distribution may be nonnormal. This gives rise to a variety of models for creating scores. Moreover, human scoring is required (even in anticipation of large-scale computer scoring). Rater scoring introduces measurement, logistic, and cost challenges.

A third issue revolves around the assumption of local independence and unidimensionality. It arises from the complexity of the underlying construct (or construct system with multiple facets) to be measured by the performance assessment. Is the assumption of local independence questionable? For example, performance assessment may involve multiple test-taker responses to information provided such that responses on several items are based on the same stimulus material. Such complexity, as is well known in reading assessment, presents challenges because factors other than a single latent ability may account for score differences. Moreover, construct complexity is likely to give rise to patterns of task- or item-level responses that exhibit multidimensionality. A single score, then, is inadequate to capture performance. Finally concern arises from the fact that performance assessments typically produce fewer item (task) scores than traditional assessments. Combining performance scores with traditional scores in an assessment system may require some form of equating and weighting so that the signal from performance assessment scores is not washed out.

And a fourth issue is that a large supply of tasks is needed, a number beyond what might be needed for a single performance assessment. This need arises because tasks are memorable, more so than a single selected-response item. Consequently new tasks need to be exchanged for old over time. The need for tasks also arises because several different forms of a performance assessment might be used in large-scale testing. The challenge is that the time and cost of constructing such tasks is considerably greater than that needed for traditional test items, whether computer-based or hands-on. Moreover their length and complexity present, at times, logistical difficulties (but computer delivery has reduced these difficulties, where appropriate/available). Finally, research on creating parallel or equivalent tasks suggests that doing so is very difficult (e.g., Bennett, 2010; Graf, Harris, Marquez, Fife, & Redman, 2010; Stecher et al., 2000).

Overview of the Report

In the five chapters that follow, we take up issues of definition; scoring, score reliability and task comparability for individual and groupwork performance assessment; and modeling and scoring of the diverse response types produced. The report concludes with recommendations for and challenges in using performance assessment for summative purposes.

Chapter II. Definition of Performance Assessment

In this chapter *performance assessment* is defined in more detail than in the section above—as an assessment activity or set of activities that requires test takers, individually or in groups, to generate products or performances in response to a complex task. These products and performance provide observable or inferable evidence of the test taker's knowledge, skills, abilities, and higher order thinking skills in an academic content domain, professional discipline, or on the job.

The definition is further clarified by five characteristics that, *taken together*, distinguish performance assessment from more traditional assessment. Performance assessments can be defined by the (a) ways they prompt test takers or groups of test takers to respond (task), (b) kinds of responses required (response), (c) way in which test takers' responses are scored (scoring), (d) accuracy with which the tasks/responses emulate a real-world context (fidelity), and (e) interconnectedness of the tasks/items within the assessment (connectedness).

Several types of performance assessment are described and examples are provided to help clarify the range of activities that are consistent with the stated definition. For each type of performance assessment, practical and psychometric considerations, including as appropriate the potential role of groupwork, are discussed.

Chapter III. Performance Assessment: Comparability of Individual-Level Scores

The use of performance-assessment tasks, as with any other type of measurement, presumes that each task is not entirely unique. Rather, the presumption is that scores on multiple, similar (parallel) tasks and their corresponding responses can be generated. Such similar tasks/responses should be substantively equivalent. Moreover, they should be statistically equivalent. That is, they should have roughly similar score means, variances, and other scale characteristics. And scores on two similar tasks should correlate highly with each other (cf. parallel-forms reliability). Finally these similar parallel tasks should correlate positively and equally with an outside criterion task. Indeed, the comparability of scores across different test forms and administration occasions fundamentally dictates the sort of inferences that can be validly drawn from assessment results. This final consideration also implies comparability in test administration, logistics, and the like.

More specifically, we can envision a performance assessment as composed of several tasks with constructed responses scored by several raters with rubrics created by several panels of experts. Sources of incomparability would include mean differences in scores over tasks, raters and panels, scoring rubrics, and training quality. Additional sources of incomparability would include interactions of these measurement facets with students.

A number of approaches to mitigating score incomparability are examined including (a) task development and form assembly, (b) anchoring and equating, (c) task and test scoring, (d) matrix sampling and reporting aggregated scores (e.g., classroom,

school), and (e) novel alternatives (reporting performance on a single task accompanied by process information without generalization to other tasks).

Chapter IV. Performance Assessment: Comparability of Groupwork Scores

Performance assessment can extend beyond the measurement of individual performance to that of individuals working in groups on a common task or set of tasks. In a real-world context, which performance assessment aims to emulate, the application of knowledge and skills to create a product or solve a problem often occurs in the company of others. More specifically, groupwork supports the measurement of some student abilities that cannot be readily measured in an individual context: (a) individuals' ability to communicate about subject matter, (b) group productivity and performance, (c) collaboration and teamwork skills, and (d) individual performance after having had an opportunity to learn or practice in groups. Moreover, the inclusion of groupwork in performance assessment positively signals its importance as a major aspect of the education standards, encouraging educators to include more groupwork in their classroom instruction.

Along with the usual challenges of measuring individual performance on complex tasks, groups working on common tasks present additional measurement challenges. Given the impact of group composition and interaction on individual performance, the interpretation of individual scores as independently reflecting individual competence may no longer be valid. Further, the same individual may be differentially impacted in different group situations and show varying performance as a result. Some ways that group functioning may differ and differentially impact results are discussed including (a) task-related interaction with others, (b) lack of involvement, (c) social-emotional processes, and (d) division of labor. Variation may also occur by (a) group composition, (b) role assignments within groups, (c) type of task, (d) specific task, (e) occasion, (f) medium for groupwork (i.e., face-to-face, online), and (g) scorers or scoring methodology.

Implications for the design of groupwork for assessment purposes are discussed. Associated issues impacting validity and reliability are examined along with their mitigations (e.g., computer simulation with student interaction with avatars, multiple group performance over randomly formed groups, aggregating individual-level data to measure class or school performance).

Chapter V. Modeling, Dimensionality, and Weighting of Performance-Task Scores

This chapter focuses on how item-level response data (e.g., item scores) can be used to produce unidimensional or multidimensional performance-assessment scores to form a measure of standing on the construct(s) being measured. The chapter is structured around three topics: (a) the models used to relate item scores to underlying constructs (concepts), (b) the particularly acute challenges around dimensionality in performance assessment scores, and (c) how different weighting strategies might be used to integrate scores from performance assessments with other components of an assessment system, recognizing that multiple skills interact in conjunctive or dissociative ways.

Models are used to link item scores to an underlying construct to form a scale and to provide a basis for estimating the magnitude of error in the score estimates. The particular challenges for performance tasks include (a) the use of trained scorers, (b) responses on a collection of different scales, (c) the desire to score processes in addition to products such as in show-your-work math items, and (d) issues of local

item dependence. Models vary in their assumptions with tradeoffs in scaling rigor. Item response theory (IRT) models make stronger assumptions than those of reliability and generalizability theories but also provide such desirable properties as interval scaling and scale equating. Indeed, IRT, with its focus on scaling, and reliability (and generalizability) theory, with its focus on measurement error, complement one another by providing different ways of evaluating performance assessment scores. Hence, models are discussed for different types of response data and different types of construct models.

The issue of dimensionality is particularly acute for performance tasks, which are inherently complex, demanding multiple abilities, skills, and actions to complete them. Multiple factor analytic models including latent-response IRT models are evaluated for addressing dimensionality and combined with exploratory models to help ferret out nuisance factors. The question of weighting becomes important when it is important to combine results from the performance assessment with multiple-choice and short-answer assessment components into one or more indicators of the overall construct of interest. Approaches and challenges are discussed for three scenarios: (a) there is a single underlying dimension; (b) there are multiple underlying dimensions, but only one relating to the construct of interest; and (c) there are multiple underlying dimensions, each related to the targeted construct.

Chapter VI. Challenges and Recommendations

At the beginning of this chapter, a set of challenges somewhat unique to performance assessment were described—the elephant in the room. The challenges were categorized into several bins of concerns including small samples of extended written or oral responses from individuals and groups in response to complex and time-consuming tasks, human scoring, multidimensionality and weighting, and comparable task generation. These and other issues are summarized, and a set of recommendations is provided for those responsible for decisions about the inclusion of performance assessment in assessment systems.

References

Baxter, G. P., & Shavelson, R. J. (1994.) Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research*, *21*, 279–298.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*, 70–91.

Collins, L. G., & Hartog, S. B. (2011). Assessment centers: A blended adult development strategy. In M. London (Ed.), *The Oxford handbook of lifelong learning* (pp. 231-250). New York, NY: Oxford University Press.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Retrieved from https://edpolicy.stanford.edu/sites/default/files/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning-report_0.pdf

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, *39*, 193–202.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*, 323–342.

Gorin, J. S., & Mislevy, R. J. (2013). *Inherent measurement challenges in the next generation science standards for both formative and summative assessment*. Retrieved from http://www.k12center.org/rsc/pdf/gorin-mislevy.pdf

Graf, E. A., Harris, K., Marquez, E., Fife, J., & Redman, M. (2010). Highlights from the cognitively based assessment of, for, and as learning (*CBAL*™) project in mathematics. *ETS Research Spotlight*, *3*, pp. 19–30.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* New York, NY: Guilford Press.

Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment, 8*, 279–315.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*(1), 41–53.

Ruiz-Primo, M. A., Furtak, E. M., Ayala, C., Yin, Y., & Shavelson, R. J. (2010). Formative assessment, motivation, and science learning. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment*. New York, NY: Routledge.

Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist, 48*(2), 73–86.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215–232.

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines.* Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/03/SmarterBalanced_Guidelines_091113.pdf

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J., & Haertel, E. (2000). The effects of content, format, and inquiry level on performance on science performance assessment scores. *Applied Measurement in Education, 13*, 139–160.

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16*, 1–26.

Wigdor, A. K., & Green, B. F. (1991). *Performance assessment for the workplace* (Vol. 1). Washington, DC: National Academy Press.

II. Definition of Performance Assessment[3]

Overview: What Is Performance Assessment? Or What Are Performance Assessments? 18
A Definition of Performance Assessment 19
Tasks 19
Responses 20
Scoring 21
Fidelity 21
Connectedness 22
What Kinds of Assessment Activities Are Not Consistent With the Definition of Performance Assessment? 22
Performance Assessment as a Guide to Rigorous Teaching, High Expectations, and Improved Test-Taker Performance 23
Types of Performance Assessment Used in Summative Student Testing Programs 23
Short Constructed-Response and Technology-Enhanced Items 23
Essay Tasks 25
Performance Tasks 26
Portfolios 27
Simulations 28
The Role of Groupwork in Each Approach to Performance Assessment 29
Features of the Definition of Performance Assessment and Approaches to Performance Assessment 30
Some Practical Considerations Regarding Performance Assessment 32
Psychometric Considerations Relevant to the Definition of Performance Assessment 33
References 34

[3] Steve Ferrara was lead author for Chapter II.

Overview: What Is Performance Assessment? Or What Are Performance Assessments?

When considering what performance assessment is or what performance assessments are, test developers, decision makers, or educators may think of performance tasks that include multiple assessment activities and multiple test-taker responses scored by trained scorers, teachers, or computers. The assessment activities in performance tasks may include a combination of multiple-choice and short constructed-response items directly related to the context of the problem presented or decision to be made with at least one (mandatory) actual context-situated performance, perhaps a culminating activity such as an essay or oral report. Special educators may think of alternate assessment portfolios, which are used to meet No Child Left Behind (NCLB) assessment requirements for test takers with significant cognitive disabilities. And experts in certification, licensure, and employment testing may think of other approaches to performance assessment such as a driver's test, a simulated role-playing task, or a simulated decision-making (in-basket) task. The aim in this chapter is to expand conceptions of performance assessment and sharpen distinctions among approaches to performance assessment.

In addressing these questions, we first considered a broad range of approaches to performance assessment:

- Short constructed-response and technology-enhanced items
- Essay prompts
- Performance tasks
- Demonstrations
- Projects
- Portfolios
- Simulations
- Learning games with feedback

In narrowing our focus, we considered our audience: psychometricians, test developers, professional educators, policy makers, and the general public who are concerned with the psychometric soundness of (a) K-12 state-testing programs that include performance assessments used for summative testing of students, whether those purposes are for public accountability reporting or high-stakes decisions; (b) performance assessments in higher education; (c) performance assessments for job-performance measurement; and (d) performance assessments for accreditation and certification.

Teachers use demonstrations, projects, and learning games with feedback for classroom assessment purposes. While much of the discussion in this and subsequent chapters applies to these approaches, we will not mention them again even though they may be embedded in large-scale assessment programs in the coming years. Consequently, we focus on approaches to performance assessment envisioned for the forseeable future to be used in large-scale testing programs:

- Short constructed-response and technology enhanced items[4]

[4] Currently, few short constructed-response and technology enhanced items meet all elements of our definition of performance assessment and the task, response, scoring, fidelity, and connectedness defining characteristics of performance assessment.

- Essay prompts
- Performance tasks
- Portfolios
- Simulations

In this chapter, we define, describe and exemplify each of these types of performance assessment,[5] all with the goal of setting the stage for addressing important psychometric considerations in performance assessment design, development, and implementation: scoring, modeling, dimensionality, weighting, comparability, and the role of groupwork.

A Definition of Performance Assessment

For the purposes of our work, we define performance assessment as

> An assessment activity or set of activities that requires test takers, individually or in groups, to generate products or performances in response to a complex task that provides observable or inferable and scorable evidence of the test taker's knowledge, skills, and abilities (KSAs) in an academic content domain, a professional discipline, or a job. Typically, performance assessments emulate a context outside of the assessment in which the KSAs ultimately will be applied; require use of complex knowledge, skills, and/or reasoning; and require application of evaluation criteria to determine levels of quality, correctness, or completeness. (Lai, Ferrara, Nichols, & Reilly, 2014)

More specifically, we define performance assessments by (a) the ways in which they prompt test takers or groups of test takers to respond (the task), (b) the kinds of responses required (the response), (c) the way in which test takers' responses are scored (scoring), and (d) the accuracy with which they represent the intended real-world situation (fidelity). In addition, (e) performance assessments often employ interconnected subtasks that form a coherent whole and that may build toward a culminating task. Until all five parts of the definition are set forth, the nature of the performance is underdetermined (e.g., Solano-Flores & Shavelson, 1997).

Tasks

Assessment tasks are directions to the test taker about what problem to solve, what product to create, or what performance or process to undertake. Directions to test takers provide information on requirements for responding, including the form or format of the response, and the features on which the response will be scored. Examples include essay tasks and activities in a performance task. The task for an assessment portfolio is likely to specify criteria for what types of products, (recorded) performances, and (recorded) demonstrations to include, how many of each, the purposes for including each type, and how they will be scored.

Performance assessment tasks vary widely. For example, tasks may require test takers to estimate the amount of fertilizer needed for a community garden and

[5] Additional details on approaches to performance assessment and examples of performance assessments are available from ETS (see https://www.ets.org/about/who/leaders), the University of Oregon (see http://pages.uoregon.edu/kscalise/taxonomy/taxonomy.html), Pearson's Research and Innovation Network (see http://paframework.csprojecthub.com/?page=home), the Stanford Center for Opportunity Policy in Education (SCOPE; see https://edpolicy.stanford.edu/category/topics/32/0?type=scope_publications), and WestEd (see http://www.wested.org/resources/preparing-for-the-common-core-using-performance-assessments-tasks-for-professional-development/).

explain why the estimation is reasonable, or conduct a research project to determine which paper towel holds the most water, or write a position paper on a current social problem, or create a collection of evidence of growth in writing argumentative essays supported by evidence and reasoning. Performance assessment tasks may require test takers to weigh multiple issues, concerns, and evidence as, for example, in conducting an investigation on qualities of objects that sink and float, or completing a research project to compare 20th century Southern Gothic writers and current authors of vampire stories. Finally, and as is evident in these examples, performance assessment tasks present complex problems that enable test takers to follow more than one solution path to complete the task and a range of ways of completing a task successfully.

Responses

Test takers' responses are intended to reflect, as closely as possible, how test takers would respond in the real-world situation represented by the assessment tasks—individually or in groups. These responses are constructed, complex, and vary in length or amount of test takers' behavior captured.

In most approaches to performance assessment, test takers must construct responses individually or together rather than select them from multiple options. For example, test takers write responses to essay tasks, perform physical actions in demonstrations (e.g., conduct a science laboratory procedure safely and successfully), prepare a research paper or science fair project, or include drawings and paintings in an art portfolio. Performance tasks may require test takers to respond to multiple-choice items and short constructed-response items in the context of completing the task (typically probing declarative or procedure knowledge underlying the overall task). As is evident in these examples, responses to many performance assessment tasks are complex in terms of understanding, processing, and responding to the task; sophistication of the knowledge and skills required to respond; and in many cases the amount of time required to complete a work product, performance, or demonstration, which could be several minutes, hours, days, or weeks; and length of the product (e.g., number of pages of an essay or written research report).

Performance assessments may require test takers to demonstrate the declarative knowledge[6] they possess as *part* of the larger assessment, and selected-response items may be more efficient for checking recall of facts, principles, and concepts. Performance assessment is considered more effective than selected-response items in assessing higher order thinking skills (e.g., Darling-Hammond & Adamson, 2010). For example, performance assessment tasks can require procedural knowledge, where test takers display their ability to apply procedures to solve problems and demonstrate complex reasoning and physical capabilities, and their scored responses represent their levels of procedural knowledge. Likewise, performance assessment can assess strategic knowledge as, for example, when test takers are required to decide which facts, concepts, and procedures to use when responding to an assessment task.

[6] Declarative knowledge is knowledge that can be stated in sentences, as in *The U.S. Civil War began in 1861* and *10 times 10 equals 100*. In shorthand, it is *knowing that*. Procedural knowledge is *knowing how*, as in demonstrating the ability to analyze historical data to discuss the causes of the U.S. Civil War and multiplying 111 x 17. Strategic knowledge, is *knowing when* to use specific facts, principles, and concepts, apply skills and procedures to solve a problem, participate in a discussion, and so forth. See Schraw, 2006, pp. 247–251 and Pressley & Harris, 2006, pp. 265–267 for formal definitions.

Scoring

A hallmark of performance assessment is that test takers construct responses that are scored for qualities like accuracy, completeness, effectiveness, justifiability, and so forth. Test-taker responses to performance assessment tasks are scored with one or more rubrics or other scoring criteria (e.g., multiple scoring rules) that are explicitly aligned with the requirements of the assessment task and required response. Scoring may involve counting numbers of acceptable responses or using a rubric to assign dichotomous scores (i.e., scores of 0 and 1). But most often scoring is based on a rubric that produces dichotomous or polytomous scores of, say, Levels 1 and 2 or highest possible scores of 3, 4, or 6 (respectively). Some scoring rubrics may have higher possible scores (e.g., score ranges of 1–12). Holistic rubrics are used to score test-taker responses on a single trait or dimension defined by broad qualities like accuracy and completeness. Multitrait rubrics are used to score responses multiple times on several qualities. Analytic rubrics may be used to parse test-taker responses into subtasks or specific steps in order to score performance on each subtask or step. In some performance tasks, responses to individual assessment activities are scored using item-specific rubrics. Determining the appropriate number of score levels in a rubric depends on the purpose of the assessment, the construct being measured, and the extent to which levels of test taker responses can be distinguished reliably and meaningfully (see Johnson, Penny, & Gordon, 2009, chapter 6, and Lane & Stone, 2006, pp. 394–395 for details on scoring rubrics.)

In large-scale, high-stakes assessments, educators or professional scorers are trained to score test-taker responses with consistency and accuracy. (See Cohen & Wollack, 2006; Johnson et al., 2009, chapters 6-8; Lane & Stone, 2006, pp. 394-400 for details on scoring rubric development, scoring procedures, and quality monitoring.) The descriptions for each score level in a rubric define features of the work product, performance, or demonstration that represent partial or complete success in achieving the accuracy, completeness, effectiveness, or other qualities called for in the assessment task.

Fidelity

Advocates of performance assessment argue that such approaches support test-taker learning because they assess test-taker knowledge and skills in real-world contexts that are relevant to test takers' experiences inside or outside of the learning environment or on a job (e.g., Shavelson, 2013). Fidelity (see Frederickson & Collins, 1989) between the assessment experience and real-world contexts often is referred to as *authenticity*. These real-world experiences may include scenarios from everyday life activities, work life activities (for older test takers), and classroom learning activities. For example, an essay task, research paper, or oral presentation task may require test takers to take a position on year-round school schedules and provide evidence to justify their position; a performance task may require test takers to observe the effects of water and a caustic solution on classroom chalk in preparation for responding to assessment activities on the effects of acid rain on headstones in a cemetery; and a simulated role play task may require nursing technician students to demonstrate proper technique in documenting a simulated patient's medical history, blood pressure, and other vital signs.

While there is consensus that fidelity is a defining characteristic of performance assessment, the important point here is related to the scenarios and assessment tasks that are intended to engage test-taker interest and effort in responding to the assessment activities. Performance assessments may be more motivating to test takers because, for example, essay tasks, performance tasks, and completing entries

for portfolios may engage test takers' interest more than decontextualized selected-response test items. However, scenarios and assessment tasks may be differentially motivating for some test takers or subgroups (e.g., Messick, 1994). On the other hand, the authentic scenarios that are intended to lend fidelity to an assessment activity also may increase reading load and other sources of complexity that may not be relevant to the construct targeted by the assessment. Furthermore, such authenticity may pose an accessibility barrier for students with disabilities, English language learners, and struggling students more generally; likewise authenticity may pose an accessibility barrier for schools or districts without the financial or human resources to provide performance assessments.

Connectedness

Performance assessments reflect a connectedness among the assessment activities and requirements for responding to those activities that contrasts sharply with the discreteness of traditional tests with multiple-choice and short constructed-response items. For example, performance tasks often require test takers to respond to several short constructed-response items that are related to the same problem to be solved, or set of stimuli to be interpreted, and the shorter responses often prepare test takers to respond to a longer, culminating assessment activity (e.g., an essay or oral report). Other performance tasks may require test takers to undertake somewhat extensive preparation (e.g., complete a hands-on science investigation of the role of light on plant growth, conduct library research) before responding at some length to a culminating assessment task. Assessment portfolios require test takers to assemble a coherent collection of evidence of work that represents the range of their accomplishments in, for example, drawing and painting or fiction writing or growth in proficiency. Even some short constructed-response and technology-enhanced items reflect connectedness by requiring test takers to develop an extended, coherent response. The connectedness characteristic of performance assessments is intended to elicit from test takers broad thinking and use of several higher order thinking skills (e.g., reasoning, critical thinking) more explicitly than is achievable with most selected-response items. The connectedness feature of performance assessment supports fidelity and enables focusing on complex, higher order thinking skills. A performance task that requires test takers to respond to several selected- and constructed-response items in preparation for to a culminating extended-response activity may require more complex reasoning and deeper thinking than is the case for responding to any one of the individual items.

What Kinds of Assessment Activities Are Not Consistent With the Definition of Performance Assessment?

Our definition of performance assessment rests on five defining characteristics that, taken together, distinguish it from more traditional tests that rely on multiple-choice and short constructed-response items. Performance assessments require products or performances that are generated in response to complex tasks; require use of complex knowledge, skills, and reasoning, and connected responses; often require multilevel scoring criteria; and emulate a context beyond the testing situation and incorporate interconnected activities. These defining characteristics rule out multiple-choice items, even those that require complex reasoning, primarily because they do not emulate real-world thinking and connected responses. Many current short constructed-response and technology-enhanced items do not meet the task, response complexity, and connectedness requirements. Performance tasks that do not meet all of the criteria (e.g., do not pose complex tasks and require complex thinking) are performance tasks in name only, but do not measure up to this definition of

performance assessment. Similarly, assessment portfolios that allow inclusion of simple performances and work that does not require test takers to complete complex tasks and produce complex responses do not align with the definition.

Performance Assessment as a Guide to Rigorous Teaching, High Expectations, and Improved Test-Taker Performance

Advocates of performance assessment argue that such assessments can act as a force for change in the content that teachers teach, how they teach that content, and their expectations for what students should know and be able to do (Darling-Hammond & Adamson, 2010). Evidence from studies in the 1990s indicates that, in fact, including performance tasks in state assessment programs is associated with positive changes in teaching and student performance (e.g., Ferrara, 2010, pp. 15–16; Lane, Parke, & Stone, 2002; Stone & Lane, 2003). Moreover, in certification examinations such as the California Bar Examination, inclusion of performance tasks has led to changes in legal education (Klein, 1983).[7] Any of these approaches to performance assessment, however, can be used to target lower level knowledge and skills (e.g., recall of facts, application of familiar algorithms). The use of performance assessment for lower level knowledge and skills alone is inconsistent with our definition of performance assessment and its role in guiding teaching and supporting student learning.

Types of Performance Assessment Used in Summative Student Testing Programs

In this section, we describe the elements of each approach to performance assessment used in summative student testing.[8] In each description, we refer to features of the definition of performance assessment and the defining characteristics. We believe that what we have to say about student testing generalizes to a large extent to other settings such as job-performance measurement.

Short Constructed-Response and Technology-Enhanced Items

Short constructed-response and technology-enhanced items may be considered part of a performance assessment as long as they are embedded in a concrete contextualized problem or decision and, along with direct observation of performance, probe the KSAs within that situation. Hence, we consider short constructed-response and technology-enhanced items in the larger context set by performance assessment.

Short constructed-response items require test takers to develop a partial or full response to a complex assessment task, as opposed to selecting a response from a set of options. These items require brief responses (e.g., no more than several minutes), in contrast to more extended responses required by essay tasks and research projects. When these items are consistent with the defining characteristics of performance assessment, they require inference, explanation, reasoning, and other complex thinking. For example, a short constructed-response item that requires students to explain that Naomi learns that her grandmother in the commissioned Grade 4 story *Grandma Ruth*[9] is loving, kind, and understanding, and to support the

[7] See also http://www.calbar.ca.gov/

[8] Additional details and real examples of each approach are available at http://pages.uoregon.edu/kscalise/taxonomy/taxonomy.html and http://paframework.csprojecthub.com/?page=home

[9] From the Smarter Balanced English Language Arts/Literacy sample items; see sample item 43600 at http://sampleitems.smarterbalanced.org/itempreview/sbac/ELA.htm The scoring rubric is available at http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/ela-rubrics/43600Rubric.pdf

explanation using descriptive details from the story, requires making inferences about characters from their words and actions and reasoning about those inferences. An item that requires students simply to describe what Grandma Ruth is like would not meet the defining characteristics of performance assessment.

Technological enhancements to short constructed-response items, such as dynamic graphing and availability of data sorting tools, may increase fidelity between the assessment activity and the targeted content standard. For example, according to the Cognitively Based Assessment of, as and for Learning (*CBAL*™) research initiative,[10] the item in Figure 2-1 from the *On the Road* Grade 8 performance task enables students to sort data in the columns to (a) locate the quartiles and median in this sample question on the use and interpret data displays learning objective, (b) interpret the summary statistics on the data set, and (c) prepare to support a subsequent recommendation on which two of 10 regions would be best for starting a delivery business.

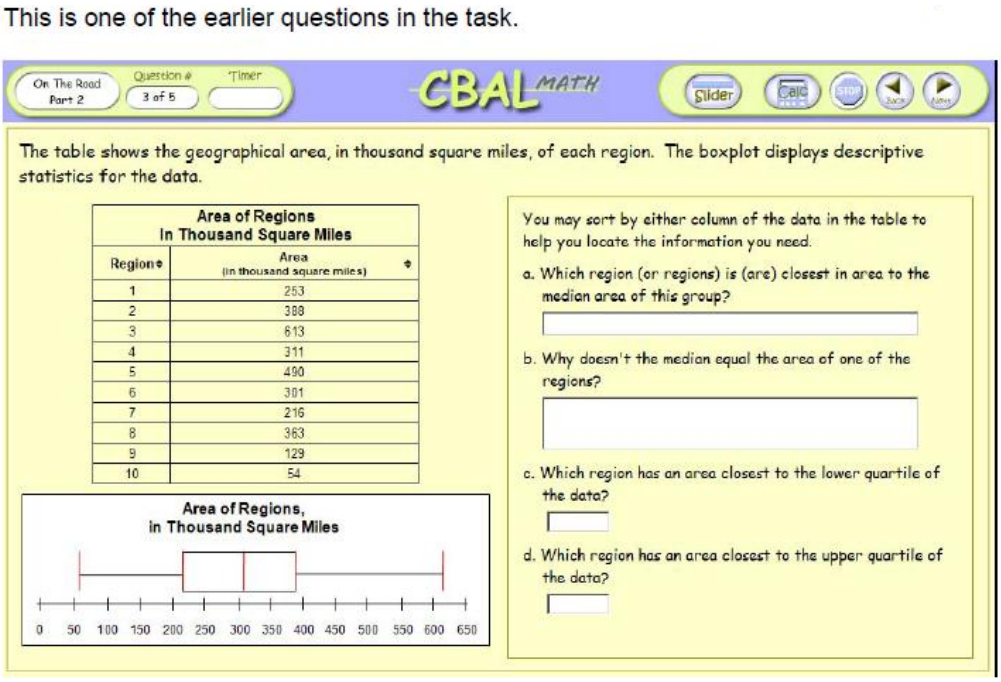

Figure 2-1. Technology-enhanced short constructed-response item from the CBAL assessment; see https://www.ets.org/Media/Home/pdf/CBALMathSampleItems.pdf for the entire task and scoring and feedback criteria. Copyright 2014 Educational Testing Service.

Short constructed-response and technology-enhanced items can require a degree of complex thinking and coherent responses. For example, dynamic concept maps (e.g., Rosen & Tager, 2014; Ruiz-Primo & Shavelson, 1996) that students can manipulate and construct on digital devices exemplify technology-enhanced performance assessment (Figure 2-2). Technological capabilities such as reorganizing, adding, labeling, and linking concepts in the map and automated, rule-based scoring enable inferences about students' understanding of complex, related concepts.

[10] See https://www.ets.org/research/topics/cbal/initiative/

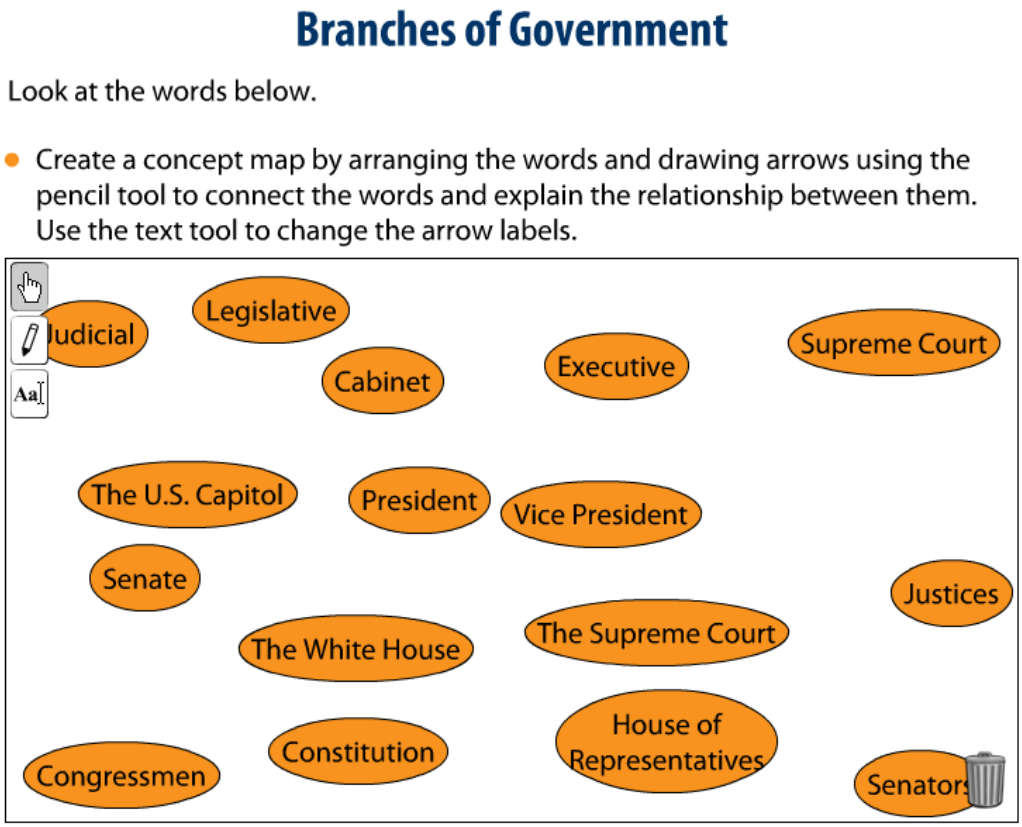

Figure 2-2. Technology-enhanced concept-map item.
See http://pages.uoregon.edu/kscalise/taxonomy/taxonomy.html, *6C. Concept Map*.

Items that simply are made more efficient by technology capabilities (i.e., short constructed-response items that do not require higher order thinking) are not consistent with our definition and defining characteristics of performance assessment. We refer to such items as technology *enabled* items. So, using drag-and-drop and hot-spot technology capabilities that require, for example, simply ordering events or ideas in a reading selection or video and that do not require, for example, identifying a claim, providing evidence, and reasoning to support the claim would be technology enablements, not technology enhancements.

Written responses to short constructed-response items typically are scored using trait rubrics that focus on correctness and relevance to the specific task that is posed. These rubrics tend to be shorter (e.g., scored as 0, 1, 2, or 3). Artificial intelligence based automated-scoring capabilities for short constructed responses are now beginning to emerge for limited types of responses (e.g., correctness of mathematical equations and short written responses; Kerr, Mousavi, & Iseli, 2013; Williamson, Xi, & Breyer, 2012, p. 3). Technology-enhanced items (e.g., concept maps) are typically scored by preprogrammed, rule-based scoring engines. On the surface, the connectedness and fidelity of short constructed-response items may be questionable. However, those items may be consistent with task and response complexity and coherence and scoring features of the definition of performance assessment.

Essay Tasks

Essay tasks require test takers to compose an extended piece of writing in response to a prompt. Although essay tasks could be categorized as a single-response

performance task, we propose it as a separate category because of its prevalence and long history in educational testing and because of the unique requirements of formulating and composing essays. Moreover, in contrast, performance tasks typically include multiple assessment activities that may include short constructed-response items, multiple-choice items, and essay prompts.

Essay prompts often identify the purpose for writing (e.g., explain a point of view, write a letter in favor of a decision with justification based on documents provided), the intended audience, and organizational or content related features of the response. Responses are lengthier than short constructed responses. They enable test takers to address a topic or purpose in depth. And they may provide test takers an opportunity to use composing processes to outline a response, write a draft, and revise to create a final response. Holistic rubrics may require an overall judgment of response quality; analytic rubrics may require separate judgments of response quality on multiple dimensions or traits (e.g., Lane & Stone, 2006, pp. 395–400). Essays can be scored by teachers or professional scorers and may be scorable by artificial intelligence based, automated-scoring engines.

Essay prompts reflect the defining characteristics of performance assessment as reflected in models of the writing process that account for the task environment, cognitive processes, long-term and working memory processes, the goal-directed nature of writing, novice and expert models of writers, and social cognitive processes (Graham, 2006, pp. 458–462). Prompts that require test takers to explain, make and support arguments, persuade, analyze, evaluate, and synthesize information from multiple sources require the higher order thinking skills that characterize performance assessment. Consider, for example the following argumentative essay task. Students review their notes taken from reading two articles and viewing a video and participate in two focused, whole-class discussions regarding the merits/limitations of school gardening programs. They then write an argumentative article for a school newsletter on starting a student gardening program at their school. The prompt indicates that writers should support their position with information from the sources and that the article will be scored on clarity of the claim on the topic, focus, and organization on the statement of purpose, elaboration of evidence, and English conventions.[11] The fidelity of essay tasks for assessment purposes arises from the role of essays in learning (e.g., undergraduate education), newspaper and magazine editorials, and analytic reports in business and professions, as well as the complex declarative, procedural, and strategic knowledge required to plan, compose, and refine a final product.

Performance Tasks

Performance tasks present test takers with a real-world-like concrete problem or situation that might be encountered in everyday school or job situations. Resources (e.g., printed, audio, visual, or manipulatives) accompany the task. They provide information needed for tackling the problem or situation. Some performance tasks may require test takers to respond to one or more subtasks that simulate the real situation and to understand and use one or more tools in responding to either a series of discrete assessment activities that are organized coherently around a common theme or a culminating activity. Or a performance task might be a single assessment prompt that requires a lengthier, more complex response such as delivering an oral presentation or conducting a science investigation. Other performance tasks may require test takers to undertake extensive preparation before responding at some

[11] From a Smarter Balanced Grade 6 sample performance task; see http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/performance-tasks/garden.pdf

length to assessment activities as in, for example, a computer simulation in which students investigate how sow bugs respond to different levels of simulated light and moisture (Shavelson, Baxter, & Pine, 1991).

Performance tasks target multiple content standards, skills, and processes. Typically, they are presented as scenarios that resemble familiar instructional activities or realistic, out-of-school activities, with at least a moderate degree of fidelity between the assessment activities and the scenario that represents the domain of generalization. They may require test takers to manipulate objects from the real-world situation (e.g., measure and mix liquids using beakers, manipulate variables in a science experiment or mathematical problem) or manipulate objects and variables in a computer simulation. Responses to performance tasks may be scored correct or incorrect (e.g., when multiple-choice items are included among other constructed-response items) or using short rubrics applied to brief responses or more complex rubrics applied to extended responses.

The assessment activities in a performance task match the defining characteristics of performance assessment when they require higher order thinking skills such as explaining reasoning, synthesizing information from multiple sources, and evaluating complex situations. For example, Partnership for Assessment of Readiness for College and Careers (PARCC) mathematical reasoning performance-based tasks call for test takers to write arguments and justifications, critique mathematical reasoning, and make precise mathematical statements. Similarly, PARCC mathematical modeling tasks call for representing real-world problems using equations and graphs. A PARCC sample Algebra II task requires test takers to distinguish linear, quadratic, and exponential models of temperature change; explain which best fits a particular range of temperature changes and why the other models do not fit well; and show work and use function notation to construct the best fitting model.[12]

Fidelity of performance tasks is achieved to the degree that the scenario that provides the context for the assessment activities mimics the real-world situation and the assessment activities and required responses elicit cognition and response processes that replicate those likely required in the real-world situation. Hands-on activities (e.g., mixing substances in beakers to conduct a science investigation, using manipulatives to reason mathematically) and group activities (e.g., peer review of written responses, group discussions) required to prepare students to make individual, independent responses often are used to provide a rationale and fidelity for responding to the assessment activities in a performance task.

Portfolios

A portfolio is a purposeful collection of a test taker's work and performances that exhibits (a) current proficiency, performance, knowledge, and skills or (b) growth over time in proficiency, performance, knowledge, and skills (Lai et al., 2014). Guidelines for assembling portfolios typically address criteria for (a) producing and selecting pieces of evidence for inclusion, (b) test-taker participation in selecting pieces of evidence for inclusion, and (c) evaluation criteria and procedures for evaluating individual pieces of evidence and making holistic judgments of the entire portfolio. Criteria may also specify test-taker self-reflection on the work exhibited in the portfolio. Pieces of evidence in a portfolio may portray both relatively routine knowledge and skills (e.g., basic academic, artistic, and other skills) and higher order declarative, procedural, and strategic knowledge. Scoring is generally completed by domain experts or trained scorers who apply rubrics or other scoring criteria and

[12] See http://www.parcconline.org/samples/mathematics/grade-7-mathematics

procedures to individual pieces of evidence and the overall portfolio. Portfolios are considered high-fidelity assessments because the pieces of evidence are intended to portray independent test-taker attainments that have been produced outside of on-demand assessment. A common concern about the use of portfolios is the degree to which peers, adults and school resources have influenced the quality of the test taker's work (e.g., Shavelson, Klein & Benjamin, 2009).

The degree to which a portfolio assessment reflects the defining characteristics of performance assessment depends on the guidelines and criteria for including individual pieces of evidence. Inclusion criteria explicitly and strategically target higher order thinking skills. For example, the Advanced Placement Studio Art assessment portfolio[13] requires students to select and submit collections of work that reflect quality (i.e., selected works), a concentration (i.e., a sustained investigation), and breadth (i.e., a range of approaches). The decisions involved in planning and producing works in each of these categories represent significant demands on declarative, procedural, and strategic knowledge.

Simulations

Simulations using digital technology overlap greatly with performance and essay tasks. Indeed, we might have combined these three categories in discussing performance tasks. However, with the potential offered by digital technology, we created a separate category. Simulations, such as the computer-simulated sow bugs performance task described above, overlap substantially with their hands-on versions but also provide an approach to performance assessment unto themselves. The goal is to simulate the criterion situation—a science experiment, an ecological environment, flying a jet fighter, military team leadership scenarios—so that the test taker's responses to learning and assessment activities can be measured, and so that their performance in the simulation can be used to predict performance in the real situation.

Simulations take on a variety forms; for example, performance tasks with scenarios (e.g., the scientist studying cooling patterns of a material; see the PARCC sample Algebra II task discussed above), computer-based simulations of environments (e.g., the EcoMUVE simulation of ponds and forests,[14] and hands-on performances for training and formative feedback (e.g., pilot training flight simulators, which are widely available via web search), and face-to-face role plays and virtual role plays (e.g., training for military officer candidates in recognizing and responding to team member behaviors).[15] Simulations are used in a wide range of learning, performance, and assessment situations, including educational achievement testing, certification and workplace testing, sports training, medical training, military training, and computer game playing. They are used to provide fidelity with the real situation, to engage test takers, to ensure practical feasibility or affordability (e.g., assessing medical diagnostic skills), and for safety (e.g., in training novice pilots to land airplanes).

The degree to which a simulation meets the characteristics of performance assessment is determined largely by the fidelity of the simulation, including the faithfulness of assessment tasks and response requirements to the real situation, the alignment of scoring criteria to evaluation criteria in the real situation, and the fidelity of the simulation itself and connectedness of the interactions and responses in the simulation. Face-to-face simulations (performance tasks) may provide the greatest

[13] See https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap-studio-arts-course-description.pdf
[14] See http://ecomuve.gse.harvard.edu/module1.html
[15] See http://www.humrro.org/corpsite/press-release/using-virtual-role-plays-develop-officer-candidates

degree of fidelity (e.g., in medical practice training, where actors simulate patients with symptoms) but with challenges due to cost and standardization across simulated patients and occasions (Clauser, Margolis, & Clauser, in press). Computer-based simulations (e.g., using avatars rather than humans) can provide cost savings (e.g., where the numbers of test takers is large) and standardization, but they can be limited in the degree to which they can simulate realistic actions, behaviors, and complexity in real environments. Simulations, if inadequately reliable (e.g., due to few responses for scoring), may not predict real-world performance as accurately as academic content domain knowledge and skill tests.

The Role of Groupwork in Each Approach to Performance Assessment

Groupwork has played a significant role in performance assessment and continues to do so. Large-scale testing programs of the 1990s in Maryland and California included group activities (National Research Council, 2010, pp. 36–37) in writing, science, and social studies performance tasks. Some essay testing programs designed in peer review and feedback on first drafts, often using specific review criteria related to the response and scoring criteria, to guide test takers prior to completing their final essay. Science performance tasks included hands-on, group science activities (e.g., the NAEP hands-on performance assessment pilot; see National Research Council, 2010, pp. 37–38; see also National Center for Education Statistics, 2012) designed to prepare test takers to respond independently to subsequent assessment activities, by activating prior knowledge and generating data for the subsequent assessment activities. Some statewide assessments with social studies performance tasks also included group discussions that were intended to prepare test takers for subsequent independent responding and to simulate debates and other discussions on social issues.

The Common Core State Standards in speaking and listening require, by definition, learners of speaking and listening skills to work with others to develop skills such as "participate effectively in a range of conversations and collaborations with diverse partners, building on others' ideas and expressing their own clearly and persuasively" (National Governors Association Center for Best Practices, & Council of Chief State School Officers, 2010, CCSS.ELA-Literacy.CCRA.SL.1) and "Present information, findings, and supporting evidence such that listeners can follow the line of reasoning and the organization, development, and style are appropriate to task, purpose, and audience" (National Governors Association Center for Best Practices, & Council of Chief State School Officers, 2010, CCSS.ELA-Literacy.CCRA.SL.4). Group assessment activities also are conducted online, as in the PISA 2018 collaborative problem-solving performance tasks, in which a test taker might collaborate with a computer agent to determine the best water and other conditions for fish in an aquarium).[16] And groupwork is playing a significant and growing role in learning games with formative feedback, as in the multi-user virtual environment (MUVE) learning game, *River City*, in which middle school students use a personal avatar to interact onscreen with computer avatars and other students' avatars to investigate the cause of illnesses in the town and share ideas and observations with teammates using a chat window.[17] Outside of educational testing, *groupwork has played a significant role in professional certification and workplace testing. For* example, in situational exercises, managerial trainees or candidates may play a role with an evaluator or another candidate, or in

[16] The PISA example is available at http://www.oecd.org/callsfortenders/Annex%20ID_PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

[17] You can explore *River City* at http://muve.gse.harvard.edu/rivercityproject/index.html

small groups, to evaluate job skills such as communication, influencing others, planning, and problem solving.

Psychometric challenges abound with group-based performance assessments, whether the focus is on individual performance, group performance, or participation dynamics (e.g., collaboration). Challenges include influence on individual and group scores of the roles that individuals take on or are assigned; status, interaction style, personality, and knowledge and skill differences in group composition; and consistency of individual behavior in different groups, for different tasks, on different occasions. These sources of variability are covered in detail in Chapter IV, and the psychometric challenges are addressed throughout this report.

Features of the Definition of Performance Assessment and Approaches to Performance Assessment

A performance assessment, as we discuss above, includes a concrete, complex task(s), responses that reflect behavior in the criterion (often real-world) situation, and a means for scoring responses. To the extent possible they should be high-fidelity simulations of the criterion situation. And the assessment typically contains a set of related subtasks. Moreover, such tasks may be carried out individually or in groups. Table 2-1 summarizes underlying features of our definition of performance assessment.

More specifically, Table 2-1 portrays general observations about the role of groupwork and independent work, hypothesized response demands, response formats, and response scoring strategies. These observations apply especially to what we refer to as *typical* design principles and practices in large-scale, summative tests, though they are relevant to classroom formative assessment practices and instruments as well. The table suggests that different approaches to performance assessment enable us to target achievement constructs and KSAs that we have not been able to target extensively or well using multiple-choice items (e.g., Ferrara & Duncan, 2011). For example, answer-explain short constructed-response items (e.g., Ferrara, Svetina, Skucha, & Murphy, 2011) that require test takers to use information from historical texts on the First Continental Congress to explain conflicts between individual rights and rule of law principles are likely to elicit evidence of conceptual understanding, reasoning, and other higher order thinking skills. Essay tasks provide a similar capability in an extended format. Empirical research indicates that the use of performance tasks with constructed-response items for targeting important constructs and yielding supportable achievement claims (e.g., Ferrara & Duncan, 2011; Shavelson, 2010; Yen & Ferrara, 1997) generally is positive. Empirical research on essay prompts is extensive and positive (e.g., Lane & Stone, 2006, pp. 398–399). Technical quality has improved since the 1990s. Koretz and colleagues (Koretz, Stecher, Klein, & McCaffrey, 1994) found significant problems with rater agreement, score reliability, and convergent and divergent validity on the Vermont writing and mathematics portfolios of the mid 1990s, though the state made significant improvements in scoring after the publication. Since then, the National Board for Professional Teaching Standards portfolio assessment has achieved high degrees of scoring accuracy and empirical support of intended inferences about accomplished teaching and, for example, the achievement of elementary school students (Goldhaber & Anthony, 2007).

Table 2-1. Features and Defining Characteristics of Performance Assessments for Each Performance Assessment Approach

Role of individual and groupwork	Hypothesized response demands	Response format(s)	Scoring
		Short constructed-response and technology-enhanced items	
Individual, independent responses	Application of a single or small number of pieces of declarative, procedural, and strategic knowledge	Written, typed, figural, or oral responses	Human scoring using short rubrics (e.g., 0, 1, 2), computer-scoring rules AI-based computer-scoring engines for some responses are emerging
		Essay tasks	
Individual, independent responses Students may collaborate on planning and critiquing a first draft	Application of a composing process plus clusters of declarative, procedural, and strategic knowledge for extended responses	A written or typed response of some length	Human scoring using rubrics (typically 1-4, 1-6) AI-based computer scoring engines for some prompts
		Performance tasks	
Individual, independent responses Students may collaborate on preparatory activities and/or when carrying out the task	Application of a small number of pieces of knowledge, conceptual understandings, and skills to multiple responses or clusters of declarative, procedural, and strategic knowledge for extended responses	Multiple written, typed, figural, or oral responses, often relatively short in length; also smaller numbers of lengthier responses	Human scoring using rubrics, computer scoring rules, AI-based scoring engines applied to responses to individual activities
		Portfolios	
Development of individual entries may be independent or may involve additional support (e.g., critique, production assistance) from teachers, mentors, and peers	Application of clusters of declarative, procedural, and strategic knowledge to select and produce portfolio entries	Written, typed, or visual work products; audio or visual recordings of demonstrations and physical performances	Human scoring using rubrics of individual portfolio entries, related collections of entries, or holistic scoring judgments of the entire portfolio

Note. Contents of cells represent *typical* large-scale testing practices.

Some Practical Considerations Regarding Performance Assessment

Some practical considerations in deciding to use a performance assessment approach are worth noting here. Performance assessments can place significant time, cost, and logistic burdens on test takers and test administrators, and both professional and automated scoring are time-consuming and costly. It is crucial that the constructs that we target with performance assessments are worth the investment and that those targets are not achievable using more efficient multiple-choice items. In addition, because performance assessment is often championed and selected for use because of its perceived influences on curriculum, teaching practice, student learning, training for licensure and certification, and the like, it is important to ensure that the expected outcomes of using a particular performance assessment approach are tightly aligned with the intended outcomes. Using logic models (e.g., Kellogg Foundation, 2004) to evaluate explicitly stated theories of action for implementing performance assessment (e.g., Dogan, 2014; McCall, 2014) can help achieve this alignment.

In addition, seemingly minor design decisions can significantly influence large matters such as classroom teaching activities and allowable inferences about student learning and capabilities. For example, science performance tasks often require students to conduct hands-on investigations, collect data, and interpret results based on the data. Because experiments can fail procedurally or yield uninterpretable results, a common testing practice is to require test takers to conduct the investigation and then direct all test takers to respond to the same "cooked" data set. However, test takers often do not recognize the relationship between the cooked data and the investigation they just completed (Stecher et al., 2000). Further, hands-on science investigations must be designed so that materials in the investigation react in expected ways, so that students are not misled or stymied by unexplainable reactions and data. And scoring rubrics should account for student responses that reflect unplanned, unexplainable results.

Richly described real-world scenarios, stimulus materials, manipulatives, and other features of performance assessments also may introduce construct-irrelevant complexity, even barriers, to students with disabilities, English language learners, students from different cultures, and other struggling students. For example, struggling readers who may know how to calculate the mean of a set of numbers may not be able to discern this relatively simple response requirement if it is embedded in a scenario on sales of items in a school store, when the data were collected, and so forth. The benefits of providing context and fidelity for most test takers can be a drawback for some. Linguistic modification (e.g., Abedi, 2006, pp. 384 ff.) and other complexity reduction strategies can help but cannot eliminate the challenges inherent in processing, understanding, and responding to performance assessments.

Groupwork is an important consideration in performance assessment. Collaboration in performance assessments is often built into assessment procedures to achieve fidelity and influence instruction in important ways, as exemplified by peer reviews of essay drafts; hands-on science investigations; group demonstrations in dance, music, and theater; project-based learning; and collaborations in learning and competitive games. In Chapter IV, we address challenges in determining how groups influence independent responses to assessment prompts, allocating credit for work and performances of individuals working in groups, and assigning scores on groupwork to individuals.

Scoring is a significant psychometric and practical consideration in performance assessment. From a psychometric point of view, scoring rubrics and other scoring criteria provide the essential link that enables inference about student knowledge and skills based on their responses to performance assessment prompts. From a practical point of view, scoring

represents significant challenges and burdens. The burdens in professional scoring include time and cost to conduct reliable, accurate scoring and the time required to return test scores. Moreover practical challenges include the substantial training and monitoring of scorers that are required to produce reliable, accurate scoring thereby minimizing error and bias. Another set of practical challenges is associated with developing exhaustive, preprogrammed scoring rules and training artificial-intelligence based, automated scoring engines (e.g., Cohen & Wollack, 2006, pp. 378-380; Shermis & Burstein, 2009).

The financial costs of performance assessment compared to multiple choice tests is of great interest and highly complex to estimate. It is clear that human scoring of complex, constructed responses to performance assessment tasks costs more than machine scoring of bubbled responses to multiple choice items—as much as 60% of the total cost of testing (Topol, Olson, & Rober, 2014, p. ii). A GAO report based on 1990–1991 data (summarized in Picus, Adamson, Montague, & Owens, 2010) estimated that the cost of tests with both performance-based and multiple-choice items would be $32 per student (in 2009 dollars); the cost of a performance assessment could be as high as $53 (again, in 2009 dollars). A recent proposed framework (Picus et al., 2010) embeds cost analyses with expected educational benefits (e.g., positive influences on teaching and learning, professional development benefits from developing and scoring performance assessments) and balances those benefits with the cost of any kind of student testing, which is less than 1% of total per pupil expenditure (p. 22). The current cost for computer based PARCC summative assessments in both ELA and math, which includes human-scored performance tasks in each content area, is $24 per student (Partnership for Assessment of Readiness for College and Careers, 2014). The cost of the Smarter Balanced end of year summative assessment, which also includes human-scored performance tasks in each content area, is $22.50 per student (Smarter Balanced Assessment Consortium, n.d.).

Psychometric Considerations Relevant to the Definition of Performance Assessment

Performance assessment presents a number of psychometric challenges. We enumerate some such challenges here and address them in subsequent chapters. One such challenge is the mélange of scores produced by subtasks in a performance assessment. For example, scored responses from short constructed-response items, technology-enhanced items, and from activities in a performance task often are scaled simultaneously with scores from multiple-choice items using item response theory (IRT) models. In other cases, composites may be formed using scaled scores from a multiple-choice test component and raw scores from an accompanying essay prompt. The challenge is to combine these scores and still retain the information provided by performance tasks.

A second challenge lies in equating scores from different performance tasks. Fewer scores typically are available from demonstrations, projects, and portfolios. Consequently scaling and equating is not likely feasible. In this case, psychometric practices can focus on scoring accuracy and consistency over occasions and managing the equivalence of assessment prompts to facilitate comparisons of performances over tasks and occasions. Attempts at statistically equating scores from essay tests (e.g., Ferrara, 1987) and scaling essay prompts (e.g., Ferrara & Walker-Bartnick, 1989) have yielded only marginal success.

A third challenge lies in the complexity—that is, multidimensionality—of performance assessments. Complex performance assessments are expected, by construct definition, to lead to scores with more than a single underlying dimension. The *connected* nature of performance assessment exacerbates this challenge in that tasks may not produce locally independent responses so some of the multidimensionality might arise from correlated errors (e.g., Yen; 1993; Ferrara, Huynh, & Michaels, 1999).

A fourth challenge lies in the wide variety of scores produced within a performance assessment. The scores may be dichotomous (i.e., correct, incorrect), multilevel (e.g., 1–3 and higher), continuous (e.g., solution speed), or some combination.

Finally, perhaps the biggest psychometric challenges are that performance assessments provide a small sample of student performance over a relatively long period of time, and student responses to performance assessment prompts are highly dependent on the prompt, response format, and topics addressed in the assessment prompt (e.g., Stecher et al., 2000). These two factors limit generalizations about students from small samples of behavior, and task sampling, expressible as person by task (p x t) variance, introduces considerable measurement error (Shavelson et al., 1993). Research on science performance tasks indicates that as many as 10–12 performance tasks may be needed to achieve dependable scores and trustworthy inferences about learners (Gao et al., 1994).

Given these challenges, one may wonder about the excitement surrounding performance assessment. For all the challenges, such assessment complements current-day assessment by broadening our capacity to fully assess constructs of interest and thereby avoid construct underrepresentation often characteristic of traditional testing. Moreover, testing in general sends a signal to educators or employers as to what is important to teach and how to teach it. Performance assessment, then, signals the importance of learning/training for doing, not just knowing.

Performance assessment, then, has an important place in assessing competence. It is incumbent upon the psychometric community to develop new models for scaling scores and evaluating reliability and validity. In an important sense, this is the psychometric challenge of the future. And the demand for alternatives is high.

References

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Mahwah, NJ: Lawrence Erlbaum Associates.

Clauser, B. E., Margolis, M. J., & Clauser, J. C. (in press). Issues in simulation-based assessment. In F. R. Drasgow (Eds.), *Technology and testing: Improving educational and psychological measurement.* New York, NY: Routledge.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and report. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–386). Westport, CT: Praeger.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Dogan. E. (2014, April). *Design and development of PARCC performance-based assessments and related research*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Ferrara, S. (1987, April). *Practical considerations in equating a direct writing assessment for high school graduation*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Ferrara, S. (2010). *The Maryland School Performance Assessment Program (MSPAP) 1991–2002: Political considerations.* Paper presented at the National Research Council workshop Best Practices for State Assessment Systems: Improving Assessment

While Revising Standards, Washington, DC. Retrieved from http://www.nap.edu/catalog.php?record_id=13013 at www.nrc.org

Ferrara, S., & Duncan, T. (2011) Comparing science achievement constructs: Targeted and achieved. *The Educational Forum*, *75*, 143–156.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Measurement*, *36*(2) 119–140.

Ferrara, S., Svetina, D., Skucha, S., & Murphy, A. (2011). Test design with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice, 30*(4), 3–15.

Ferrara, S., & Walker-Bartnick, L. (1989, April). *Constructing an essay prompt bank using the Partial Credit model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7,* 323–342.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National Board certification as a signal of effective teaching. *The Review of Economics and Statistics*, *89*(1), 134–150.

Graham, S. (2006). Writing. In P. A. Alexander & P. H. Winne (Eds), *Handbook of educational psychology* (2nd ed., pp. 457–478). Mahwah, NJ. Lawrence Erlbaum Associates.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: Guilford Press.

Kellogg Foundation. (2004). *Logic model development guide*. Available from http://www.smartgivers.org/uploads/logicmodelguidepdf.pdf

Kerr, D., Mousavi, H., Iseli, M. R. (2013). *Automatically scoring short essay for content* (CRESST Report 836). Los Angeles, CA: University of California, National Center for Reearch on Evaluation, Standards, and Student Testing (CRESST).

Klein, S. P. (1983). *Measuring research skills on a bar examination.* Santa Monica, CA: The RAND Corporation.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, B. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*(3), 5–16.

Lai, E. R., Ferrara, S., Nichols, P., & Reilly, A. (2014). *The once and future legacy of performance assessment*. Manuscript submitted for publication.

Lane, S., Parke, C. S., & Stone, C.A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment, 8*(4), 279–315.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: Praeger.

McCall, M. (2014, April). *Multi-state consortium [Smarter Balanced] performance assessment: Theory into action.* Presentation at the annual meeting of the National Council on Measurement in Education, Philadelphia.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

National Center for Education Statistics. (2012). *The nation's report card: Science in action: Hands-on and interactive computer tasks from the 2009 science assessment* (NCES 2012-468). Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2009/2012468.pdf

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *English language arts standards, anchor standards, college and career readiness anchor standards for speaking and listening*. Retrieved from http://www.corestandards.org/ELA-Literacy/CCRA/SL/

National Research Council. (2010). *State assessment systems: Exploring best practices and innovations: Summary of two workshops.* Washington, DC: The National Academies Press.

Partnership for Assessment of Readiness for College and Careers. (2014, May 2). *States select contract to help develop and implement PARCC tests* [Press release]. Retrieved from http://parcconline.org/states-select-contractor-help-develop-and-implement-parcc-tests

Picus, L. O., Adamson, F., Montague, W., & Owens, M. (2010). *A new conceptual framework for analyzing the costs of performance assessment.* Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Pressley, M., & Harris, K. R. (2006). Cognitive strategies instruction: From basic research to classroom instruction. In P. A. Alexander & P. H. Winne (Eds), *Handbook of educational psychology* (2nd ed., pp. 265–286). Mahwah, NJ. Lawrence Erlbaum Associates.

Rosen, Y., & Tager, M. (2014). Making student thinking visible through a concept map in computer-based assessment of critical thinking. *Journal of Educational Computing Research*, *50*(2), 249–270.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996) Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, *33*(6), 569–600.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment*, *4*(6). Available from http://www.jtla.org

Schraw, G. (2006). Knowledge: Structures and processes. In P. A. Alexander & P. H. Winne (Eds), *Handbook of educational psychology* (2nd ed., pp. 245–263). Mahwah, NJ. Lawrence Erlbaum Associates.

Shavelson, R. J. (2010). *Measuring college learning responsibly: Accountability in a new era.* Stanford, CA: Stanford University Press.

Shavelson, R. J. (2013). An approach to testing and modeling competence. *Educational Psychologist, 48,* 73–86.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 41–53.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, *4*, 347–362.

Shavelson, R. J., Klein, S., & Benjamin, R. (2009, October 16). The limitations of portfolios. *Inside Higher Education.* Retrieved from http://www.insidehighered.com/views/2009/10/16/shavelson

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E.W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36,* 61–71.

Shermis, M. D., & Burstein, J. (Eds.). (2009). *Handbook of automated essay evaluation*. New York, NY: Routledge.

Smarter Balanced Assessment Consortium. (n.d.). *Frequently asked questions*. Retrieved from http://www.smarterbalanced.org/resources-events/faqs/

Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice*, *16*(3), 16–24.

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J., & Haertel, E. (2000). The effects of content, format, and inquiry level on performance on science performance assessment scores. *Applied Measurement in Education*, *13*(2), 139–160.

Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16*(1), 1–26.

Topol, B., Olson, J., & Roeber, E. (2014, February). *Pricing study: Machine scoring of student essays*. Retrieved from http://www.assessmentgroup.org/uploads/ASAP_Pricing_Study_Final.pdf

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

Yen, W. M., & Ferrara, S. (1997). The Maryland School Performance Assessment Program: Performance assessments with psychometric quality suitable for high-stakes usage. *Educational and Psychological Measurement*, *57*(1), 60–84.

III. Performance Assessment: Comparability of Individual-Level Scores[18]

Introduction 39
Substantive Equivalence 40
Statistical Equivalence 41
Challenges to Score Comparability 41
  Score Variability Due to Tasks 42
  Score Variability Due to Raters 42
  Score Variability Due to Occasion 43
Mitigating Challenges to Score Incomparability 44
  Task Development and Form Assembly 44
  Test Equating 45
    Common Item Equating of Performance Assessments 47
    Equivalent Groups Equating of Performance Assessments 48
    Single-Group Equating of Performance Assessments 49
  Task and Test Scoring 50
  Computerized Scoring of Performance Assessments 50
  Matrix Sampling of Tasks 51
Conclusions 51
References 53

[18] Tim Davey was lead author for Chapter III.

Introduction

Tests differ widely in the sorts of inferences that their scores are intended to support. For example, consider a test that is administered on a single occasion each spring to determine which of a nation's students are admitted to college the following fall. Assuming that a single version, or *form* of the test is administered, the performance of any test taker can be fairly compared to that of any other. The percentile rank of any test taker could also be determined within the cohort of test takers who tested on a given form. Further, the performance of any test taker could be compared to some fixed performance threshold, provided that threshold was established on the same test form administered to the test takers being compared.

Other testing programs are more ambitious in their design, deploying multiple test forms across multiple administration occasions each year. The main reason for use of multiple test forms is simple: A single form will not remain secure under repeated administration in an environment where stakes are attached to score use. The price paid for increased security is that score comparability is no longer transparently assured but instead can be achieved only by considerable effort. Strictly and ideally defined, scores are truly comparable when each test taker would achieve the same score regardless of the form that test taker is administered (and the circumstances of that administration). Unhappily, this vision of comparability can only be approximated in practice. Other, weaker definitions of score comparability are therefore offered below. Determining the extent to which scores are comparable is critical because it fundamentally dictates the sort of inferences that can be validly drawn from assessment results. Examples of inferences that can be affected by score comparability include:

- Mary performed at the 87th percentile of all 4th graders this year. (Was Mary's score comparable to those of other 4th grade students?).
- John's mathematics proficiency grew at an average rate over the course of 6th grade. (Were John's scores at the beginning and end of 6th grade comparable?).
- Sixty-two percent of third-year medical students at Northern University were judged as proficient relative to their state's standards for clinical performance. (Were Northern University scores comparable to those on the test form for which the performance standards were set?)

Of course, all of these inferences depend on much more than score comparability, but each is valid only to the extent that scores produced by different test takers in different places at different times on different test forms are directly comparable. Although scores can remain useful lacking comparability, the sorts of inferences that can then be supported are correspondingly limited. The example above of a test administered once each year is illustrative. Test takers administered the same form could be compared with one another but, without comparability, not to test takers tested with a different form. Similarly, test takers could be compared to performance standards set on the form they were administered but not to standards set on other forms that may or may not produce comparable scores.

Considerable methodological and procedural machinery has been developed over the past century to measure and ensure score comparability (Holland & Dorans, 2006). A consensus of this work is that there are two requirements for comparability. The first is that the tests or test forms being compared are substantively equivalent, meaning they measure the same thing (often called the *construct*). The second is that the test forms and their scores have equivalent statistical properties (e.g., produce the same distributions of scores, measure with the same level of reliability, and correlate equally with other tests or criterion measures). The means for ensuring substantive equivalence, the mechanisms for achieving

statistical equivalence, and the methods for checking whether each holds are well understood and generally effective in the context of traditional multiple-choice assessments (Kolen & Brennan, 2004). However, both substantive and statistical equivalence are inherently more difficult to achieve with performance assessments.

Substantive Equivalence

With both multiple-choice and performance assessments, the items or tasks from which test forms are built serve as the concrete representation of the construct being measured. Each test form is (or at least should be) built with respect to a blueprint that simultaneously guides test assembly by specifying the nature and number of the tasks that a form comprises and serves to fix the construct measured across forms. It may be helpful to conceive of a universe that includes all of the tasks considered appropriate for inclusion on a given test, with any single form viewed as a sampling of that universe. One goal of test assembly is then to ensure that the tasks selected for each form are a representative sample of the universe. Unbiased sampling of the task universe can be facilitated (but not guaranteed) by stratification, which divides the task universe into subdomains and then samples separately from each. Stratification is essentially the role of the blueprint, which partitions the task universe and specifies the density and depth of the sampling from each partition.

Stratification is not the only way of promoting broad and representative sampling of the task universe. Consider, for example, simply increasing the size of the item or task sample. The practical advantage of building a test from many items lies in the premise that the interactions between test takers and tasks can never be fully understood or explained. Although many of the important task characteristics that drive test-taker performance can be identified, formalized in the blueprint, and used to stratify the task sample, myriad minor, subtle. or unexpected influences on performance may remain hidden. Building the test from many, small independent tasks can allow these minor influences to be randomized over the sample of tasks so that none appears with sufficient frequency or prominence to induce bias. This premise appears in both sampling and in experimental design; the common goal is to control those factors of the sample or experiment that can be controlled and to minimize the impact of the remaining factors that cannot be controlled by randomization.

Larger task samples are usually more easily recommended than implemented. The traditional path to larger samples is to reduce the size of the sample unit by using individual, discrete items rather than larger, more context-laden tasks.[19] However, performance assessments include large tasks not by choice but rather because measurement of the construct demands them. Altering, or redefining a construct to one that will submit to being measured by small, discrete items threatens test validity both from a content perspective and in the sense that appropriate feedback is returned to the educational system. The alternative of building a long test composed of many, large, context-laden units increases both the cost of test development and the time required for administration, neither of which may be desirable to a test's stakeholders.

[19] Task size can be viewed several ways. As described in the previous chapter, many performance tasks include an extensive and detailed context, within which a complex problem must be solved. This can require a good deal of work on the part of the test taker, both to absorb the context and to generate and communicate a solution. All of this requires time, meaning performance assessments generally yield far less measurement information per unit time than do traditional selected-response items. To improve efficiency, some performance tasks build multiple problems into a single context, the reasoning being that the incremental time required to collect the responses to these additional problems is less than what would be required to establish a new context for each. The result is that tasks become larger and more time-consuming. A test composed of more than a handful of such tasks might, therefore, require a prohibitively long time to administer.

Test developers may also be tempted to narrow the definition of the construct being measured (and therefore the breadth of the task universe required to measure it), on the theory that a narrower task universe is more readily sampled than a broader one. However, doing so can threaten content, predictive, and construct validity (e.g., Cronbach, 1990).

The nature of performance assessments, then, often makes the collection of tasks selected for each test form a coarser, often less representative, sample of the universe of tasks. Methods for evaluating whether different tests or test forms measure the same construct are of two sorts. Judgmental approaches rely on the insight and experience of subject matter experts who review each form and gauge whether the forms are measuring the knowledge domains in the same ways with the same emphases (Kane, 2006). Empirical approaches are based on analyses of the relationships both between a test's task scores and between those scores and scores on various outside criteria. Factor analysis, multidimensional item response models, and structural equations models are the common tools (Bartholomew & Knott, 1999; see Chapter V). Because each of these tools is difficult to apply, hard to interpret, and potentially fallible, it is best to apply several and seek converging evidence.

Statistical Equivalence

Holland and Dorans (2006) defined the scores produced by alternate forms of a test to be comparable when the forms meet three conditions. The first two conditions are the usual requirements that the forms are equally reliable and produce equivalent score distributions when administered to equivalent test-taker samples. Their third condition requires the test forms to have the same pattern of correlations with external criterion measures (e.g., other test scores or outcome measures). This last requirement empirically assesses substantive equivalence, which is important because scores produced by tests of very different constructs can be easily manipulated to have comparable reliabilities and score distributions (van der Linden, 2000). However, these tests should fail the third requirement by having different correlations across a range of external criterion measures.

External correlations notwithstanding, statistical equivalence is much more easily measured than substantive equivalence, at least in theory. All that is needed is to administer the test forms to equivalent groups and compute reliabilities and score distributions. Some testing programs can do this without difficulty because they routinely administer alternate test forms to equivalent test-taker groups in the normal course of their score equating process. But other programs make use of equating designs that do not require equivalent groups, meaning such samples could be drawn only with effort and expense. These programs may rarely (or never) revisit the question of score comparability, assuming it was established in the first place.

Challenges to Score Comparability

Following our definition (Chapter II), we conceive of a performance assessment as composed of a sample of tasks (which produce selected- and extended constructed-responses) that are administered on one or more occasions and then scored by one or more raters. Variability in scores due to true differences in performance across test takers constitutes the signal we wish to detect. However, variability due to tasks and raters, and their interactions with test takers, constitute noise that degrades score reliability and, more importantly, decreases the extent to which scores are comparable across test takers. Important sources of such noise are identified and briefly discussed below. These sources fall generally into three categories: those due to the tasks themselves, those due to the raters and the rating process, and those due to the occasions across which test takers test.

Score Variability Due to Tasks

Task-sampling variability refers to both variation in (a) task difficulty (mean differences among task scores), and (b) test takers' success on one or another task—that is, test taker by task interaction. Research shows large variability in individuals' scores across tasks, the consequence of which is generally low reliability of scores produced from a single administration of a small number of tasks.

Examples of large task variability come from a variety of performance assessment domains. Generalizability studies of performance-based assessments in science, mathematics, medicine, and military specialties show two effects related to task variability. First is a non-negligible main effect for tasks, meaning some tasks are easier than others for all test takers. This is not unexpected because cognitive tasks of all sorts – both constructed and selected response– are subject to variation in difficulty due to a wide variety of factors. The second finding is more troubling, this being a large interaction between test takers and tasks. This indicates that the tasks that some test takers find difficult are not the same tasks that others find difficult. The test taker by task interaction effects can be so large that achieving moderately dependable measurement (e.g., reliability estimates of .80) may require upwards of 10 to 24 mathematics, science, medicine, or military tasks (Lane, Liu, Ankenmann, & Stone, 1996; Richter-Lagha, Boscardin, May, & Fung, 2012; Shavelson, Baxter, & Gao, 1993; Shavelson, Mayberry, Li, & Webb, 1990; Webb, Shavelson, Kim, & Chen, 1989).

Score Variability Due to Raters

Rater-sampling variability results from disagreement among raters in the scores assigned to the same responses. Disagreements can arise because raters approach their work with differing degrees of stringency or leniency. This leads some raters assigning higher scores, on average, to a collection of responses than those scores assigned by other raters. Raters may also differently interpret the scoring *rubric*, or the instructions intended to guide scoring. Raters may also drift over time in the scores they assign, grading responses more leniently at some times and more stringently at other times. Finally, raters may disagree more idiosyncratically, by differing in the scores they assign to particular test takers (thereby constituting test taker by rater interaction).

The scale on which performances are rated can also affect score variability. For example, writing samples may be scored by a single holistic judgment of writing quality (often called a holistic rating). Alternatively, raters may evaluate each feature of the writing (e.g., content, organization, mechanics). This is often called *analytic scoring*. Ratings may also extend beyond technical writing quality to encompass substantive factors such as the clarity or persuasiveness of the writer's position. A single writing sample may even be scored through a combination of these scales (Lane & Stone, 2006). The scales may also differ in the number of score points on which performance is rated. Examples include a 4-point scale for scoring writing dimensions of development, organization, attention to audience, and language (Maryland State Department of Education, 1996), a 5-point scale for scoring mathematics dimensions of mathematical knowledge, strategic knowledge, and communication (Lane, 1993), or a 7-point scale for judging grammar in foreign language speaking (Bachman, Lynch, & Mason, 1995).

Research on the impact of scale characteristics on rater performance is, thus far, inconclusive. For example, when rating performance on hands-on science performance tasks, Klein et al. (1998) found high agreement between analytic scores (sums of scores on specific items such as constructing tabular results when using magnets to sort metallic and nonmetallic trash) and holistic scores (the overall quality of the student's performance on three standards, conceptual understanding, performance, and application). Conversely,

Moss, Cole, and Khampalikit (1982) found low agreement between analytic and holistic ratings of written language skills, specifically analytic scores based on the number of spelling, capitalization, punctuation, and expression errors, and holistic scores based on judgments about organization and clarity. Lane and Stone (2006) suggested that agreement between rating types may depend on similarity of the criteria reflected in the scales.

Score variation due to rater effects of all sorts tends to be small if raters are adequately trained and monitored (Shavelson et al., 1993). Both human raters and computerized scoring systems can and do produce consistent (reliable) scores (Shavelson, 2010; Shermis & Burstein, 2003; Steedle & Elliot, 2012). That said, less well trained raters may produce substantial and problematic errors (e.g., McCaffrey, Yuan, Savitsky, Lockwood, & Edelen, 2014). Moreover, human raters generally do need to be monitored to insure that they stay calibrated and free of score drift. Although computerized scoring systems may appear immune to the problem of inconsistent standards across time, changes and enhancements to the scoring software can certainly introduce variation much akin to score drift (Williamson, Xi, & Breyer, 2012).

Because rating of constructed responses is resource intensive, less costly and time-consuming scoring paradigms are actively sought. Computerized scoring is one such possibility. However, as detailed below, its applicability is currently limited to certain constructs and types of tasks. Another option is to have test takers rate their own work (self-assessment) or the work performed by their peers (peer assessment). Evidence about the validity of such ratings is mixed. Reviews of research comparing peer and teacher assessments of a wide variety of activities (e.g., research proposals, posters summarizing research, biochemical laboratory reports, oral reports in pharmacology, psychology term papers, psychology debates, essays about literature) in a range of fields (e.g., science, law, business, engineering, psychology, management, medicine) have reported a wide range of correlations (e.g., Falchikov & Goldfinch, 2000; Topping, 1998; van Zundert, Sluijsmans, & van Merrienboer, 2010). Those reviews suggested that agreement between peer and teacher ratings may be higher when peers make global judgments than when they rate separate dimensions of performance. As expected, training can improve students' skills in performing peer assessment (e.g., reducing discrepancies between student and instructor ratings of students' projects, Liu & Li, 2014).

Score Variability Due to Occasion

Test-taker performance on the same tasks can vary across time both systematically and idiosyncratically. Systematic differences are revealed by mean score differences across occasions and can be due to disturbance, memory, or familiarity effects. Idiosyncratic differences are present when some test takers perform better on one occasion while others perform better on a different occasion. This effect is measured as a test taker by occasion interaction. Evidence of the volatility of student performance over occasions is widely reported (e.g., Ruiz-Primo, Baxter, & Shavelson, 1993). Generalizability studies of performance assessments administered on multiple occasions have typically shown small differences in task score means across occasions but sometimes found large test taker-by-occasion-by-task interactions (Shavelson et al., 1993). In mathematics and science performance assessments, how a particular student performs on a task depends both on the particular task and the occasion on which it is administered (e.g., McBee & Barnes, 1998; Shavelson, Ruiz-Primo, & Wiley, 1999; Webb, Schlackman, & Sugrue, 2000). Score comparability is therefore degraded not just by variability in task sampling across test forms but rather by interactions between task variability and the sampling of the occasions of test administration. An intriguing, albeit troublesome, explanation for this volatility in scores over time is that test takers may change their approach for solving a particular task from one occasion to the next (Ruiz-Primo et al., 1993).

The variability in scores due to test taker-by-occasion interactions may call into question the utility of stratifying performance tasks into homogeneous subsets as a cost-effective strategy for improving reliability (Shavelson et al., 1999). Because tests are generally administered just once to a given sample of students, measurement error due to sampling of tasks and sampling of occasions is inextricably confounded (Cronbach, Linn, Brennan, & Haertel, 1997). Stratification may be effective at reducing the impact of task variability itself, but not at reducing the impact of occasion variability. In their analysis of the variability of science performance assessment scores, Shavelson et al. (1999) showed that stratifying tasks had only a minimal impact on reducing the combined effect of task and occasion variability and, consequently, had little effect on the number of tasks needed for dependable measurement.

Mitigating Challenges to Score Incomparability

By their nature, performance assessments have been shown to pose challenging risks to the substantive and statistical comparability of test scores across test takers and test forms. The following section accordingly presents a number of strategies and methods for coping with and mitigating these risks to the extent possible. Risks can be addressed during the task development and form assembly processes through the methods used for anchoring and equating and possibly during the test scoring process itself.

Task Development and Form Assembly

Task variability across test forms might be controllable by developing detailed test blueprints and adhering to them rigidly during test assembly. Success would depend on identifying and managing as many influences on task performance as possible. Types of knowledge and understanding, level of inquiry, types and sequences of actions for students to perform, are all potential (and perhaps essential) candidates for control. However, even doing so does not guarantee task or score comparability. For example, Stecher et al. (2000; see also Solano-Flores & Shavelson, 1997) used a detailed shell (blueprint) to construct two forms of a science assessment with conceptually similar performance tasks. The shell specified the core content, the stages of the scientific investigations to be carried out, the levels of inquiry, and the task formats. Despite the attempt to create highly similar test forms, students' performance on tasks designed to be similar in format, content, and inquiry level was no more alike than their performance on tasks that varied along these dimensions.

At the limit, new test forms could be built from tasks cloned from or modeled on those used on some initial or base test form. The intent is to use task modeling to produce parallel forms composed of identically performing tasks. In this case, the collection of task models essentially serves as the form blueprint. This strategy may prove effective with simpler tasks but would be increasingly difficult to implement as tasks became more complex. There is also the concern that use of cloned tasks may lead to a narrowing of the measured performance domain and can promote teaching tactics that focus on the specific elements measured by the cloned tasks rather than on the subject matter more generally.

Shavelson and colleagues used a cloning approach to develop versions of science tasks with parallel structures and similar appearances (Baxter & Shavelson, 1994; Solano-Flores & Shavelson, 1997). *Electric Mysteries*, for example, required students to test six mystery boxes to determine their contents (two batteries, a wire, a bulb, a battery and a bulb, or nothing) and *Bugs* asked students to carry out three experiments to determine whether sow bugs choose light or dark, damp or dry, or some combination of these environments. Student performance was quite consistent across the six electric circuit tasks, but less consistent across Bugs, composed as it was of only three subtasks. Some writing tests have also achieved success by building multiple essay prompts from the same task model. They

administer the prompts in equivalent groups and use statistical estimates of the performance characteristics of each prompt to select parallel forms for operational use.

In most cases, the surest way of limiting form variability is by increasing the number of tasks on a form. Because doing so may unduly extend testing time, one option is to require each test taker to complete different sets of tasks. In a science assessment, students may design a biology experiment, critique a physics experiment, and interpret geology and astronomy datasets rather than carrying out an entire experiment in one area (Kane, Crooks, & Cohen, 1999). Kane et al. (1999) also suggested abbreviating certain steps in a task (especially routine ones) to reduce the time required.

A more common solution is to use what can be termed a mixed-format test (Kim & Kolen, 2006). Mixed-format tests typically comprise a few large constructed-response tasks and a larger number of smaller, independent selected-response items related to the overall task. A mixed-format test can then be seen as a compromise that uses a combination of small, discrete items and larger performance tasks, with each type used to measures those aspects of the construct to which they are best suited. As described below, accompanying performance tasks with selected-response items can also greatly facilitate use of test equating methods as a means of improving score comparability.

Test Equating

Test equating methods are traditionally employed to make data-driven adjustments to scores to ensure comparability across test forms (Kolen & Brennen, 2004). Score adjustments can take place either before or after test administration, with both approaches requiring that data be collected according to one of three basic designs, illustrated in Figure 3-1.

Under the *single-group* design, all test takers are administered both test forms being equated. The *counter-balanced* design is an important variant of the single-group design, intended to correct biases inherent in the order of form presentation to each test-taker. For example, working through Form A might leave test takers tired or disinterested, leading to lower performance levels on Form B. Conversely, Form A may serve as a warm-up, resulting in better performance on Form B. The counter-balance solution is to split the total sample in two and administer Form A first in one half and Form B first in the other.

In the *random-* or *equivalent-groups* design, each form is administered to distinct, but statistically equivalent samples.

Finally, under the *common-item design* the test forms are administered to nonequivalent samples but share a set of items in common. These common items are usually called *anchor items*. Two variants of the common-item design can also be distinguished. Under the *internal anchor* design, the anchor items are contained within the test form and contribute to the score being equated. An *external anchor,* in contrast, is a collection of items that stands outside of the test form and that do not contribute to the score being equated.

Each of the three basic equating designs (and their variants) has both benefits and drawbacks. The single-group design (particularly in its counter-balanced version) is statistically strong, capable of producing accurate equating results with relatively small-taker samples. However, the requirement that each test taker take both test forms is operationally inconvenient. As such, it is rarely implemented in practice outside of special circumstances[20] (Livingston, 2014).

[20] One of these circumstances involves equating a shortened version of test to the longer version of itself. This situation arises when one or several items have been removed from a test form that is reused after having been

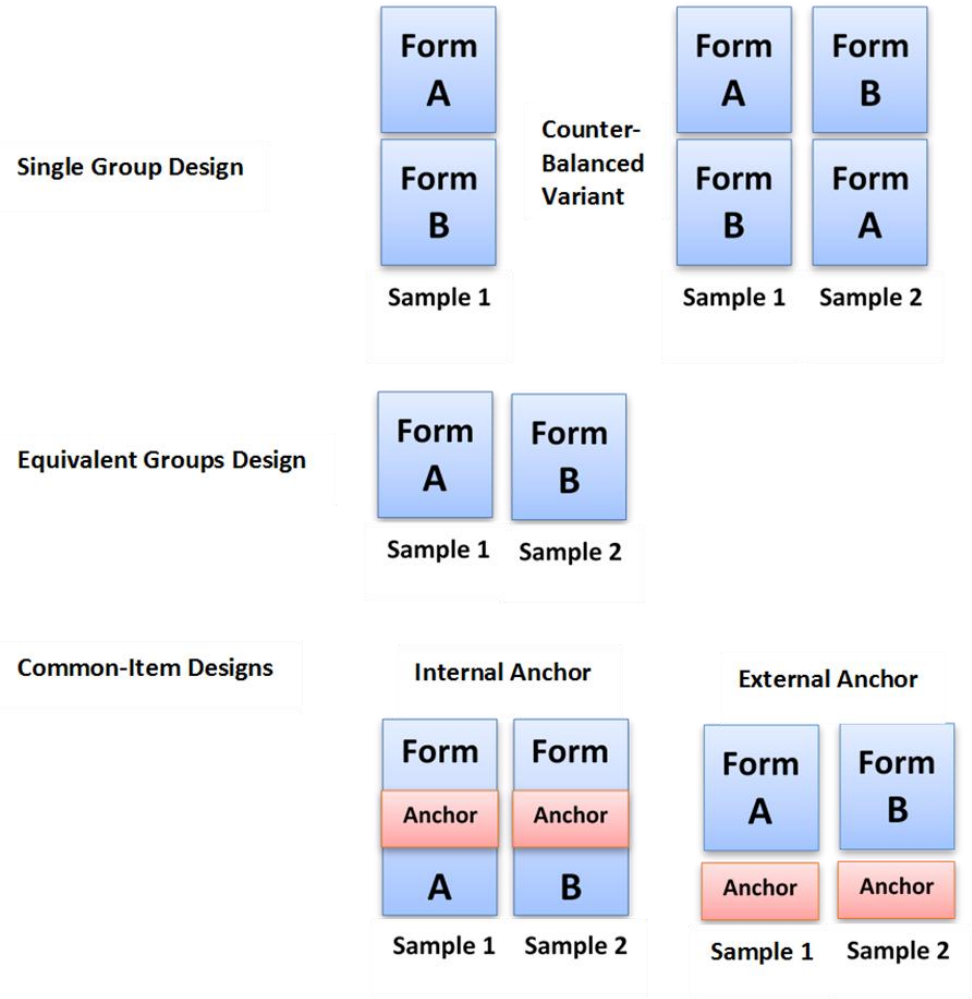

Figure 3-1. Basic equating designs.

The equivalent groups design is weaker than the single-group design, allowing the relative difficulty of the two test forms to be compared only to the extent that the test-taker samples are in fact equivalent. Because the chances of random equivalence improves as sample sizes increase, the sample requirements are greater than under the single-group design. The compensation for this is that each test taker is required to take only one test form. Success depends in large part on the extent to which the equating relationship estimated from the equivalent groups generalizes to the full test-taker population.

The common-items designs gauge the relative difficulty of the two test forms not directly but rather through reference to the anchor items common to both forms. Comparing the performances of the two test-taker groups on these items allows one to determine whether one group is more able than the other. This, in turn, permits the relative difficulty of the test forms to be established. Common-item equating requires more complex (and so riskier)

previously equated. Because the members of the equating sample took both the full and shortened versions of the test, the single-group design applies.

methodology than the other, more straightforward designs. However, it compensates for this through greater operational flexibility and convenience.

Common Item Equating of Performance Assessments

To serve properly as anchors, the sets of common items shared by test forms must both accurately represent the construct and allow for sufficiently reliable measurement. It is therefore usually recommended that anchor-item sets follow the same blueprint that drives form assembly and that they constitute a quarter or more of the test form. These are relatively easy conditions to meet with tests composed of large numbers of independent, discrete items. However, anchors are hard to extract from performance assessments consisting of only a handful of tasks. Any subset of these tasks that incorporates less than the bulk of the test form is likely to be either an unrepresentative sample of the construct or provide insufficiently strong measurement. Because constructed-response tasks can be quite memorable to test takers, security concerns may prevent their use as equating anchors, where they would need to be administered on multiple occasions.

A possible solution to both problems identified above is to use external anchors. As described above, an external anchor is a set of items appended to or embedded within a test form that do not contribute to test-taker scores but rather serve only the purpose of equating. The first advantage of an external anchor is that it can be arbitrarily large, solving any problems regarding content representativeness. Of course, very large anchors can add significantly to testing time without perceived value. However, it is possible to split a large external anchor into several smaller, less time-consuming pieces, only one of which is administered to any given test taker, as long as the construct measured is preserved in each. Splitting the anchor across test takers also limits the exposure of a single task to a subset of the test-taker population, possibly enhancing security.

Whether based on internal or external anchors, the common-item design depends on the anchor items or tasks performing equivalently across test forms and administration occasions. Several characteristics specific to performance assessments can put this requirement at risk. First, anchor tasks may perform differently across occasions due to rater inconsistency. Although proper training and monitoring of raters can promote scoring consistency, they cannot ensure it. For example, Fitzpatrick, Ercikan, Yen, and Ferrara (1998) found considerable inconsistency in scores assigned to students' responses on the Maryland School Performance Assessment Program's open-ended, constructed-response items by two groups of raters carrying out their judgments one year apart, especially in language arts areas. The groups of raters used in different years differed in severity. Kim, Walker, and McHale (2010) also observed that attention to rating consistency across occasions is essential to satisfactory use of constructed-response anchors. Rater inconsistency across occasions can even result in an anchor that includes an appropriate mix of selected- and constructed-response items performing worse than one consisting of selected-response items alone. Tate (2000) recommended checking consistency of raters across scoring occasions by having current raters rescore responses previously scored by an earlier cohort of raters. This is sometimes called comparability or trend scoring (Kim et al., 2010). Such checks allow differences in rating standards to be isolated from other factors that can change task performance over time.

One of these other factors is a real change in how test takers perform on the anchor items. As noted above (e.g., Ruiz-Primo et al., 1993), test takers may not respond identically even when they encounter the same performance tasks on different occasions, possibly because their approach to completing the task changes (e.g., changing their experimental design when testing sowbugs' preferences for lighting and moisture in their environment on a science assessment task). This inconsistency complicates the ability to draw inferences from a single administration of anchor tasks for any group of test takers.

A second other factor arises from the use of anchor items on several occasions. Because anchor items must be used on several occasions, inadvertent disclosure is a possibility. Test takers with prior exposure to the anchor items will almost surely perform better than test takers seeing the items for the first time.

A novel possibility for overcoming the challenges posed by performance assessment is to use anchor tasks that appear to be different but that perform identically. This can circumvent the security problems associated with administering the same tasks across multiple occasions. However, controlling task performance to the extent that different sets of tasks can effectively serve as common-item anchors is extraordinarily difficult. Success would depend on being able to generate, from a model, tasks that had equivalent performance characteristics. Taken to its extreme, task modeling could, in theory, make equating superfluous by producing alternate test forms that perform so similarly that equating is unnecessary. Even if identical form performance proves unachievable, rigidly controlling the difficulty of each assembled test form remains an important goal since, in all cases, equating works best when it is needed least.

Mixed-format tests built of both constructed- and selected-response items offer other, often highly practical, anchoring possibilities. The first is to equate via an anchor consisting solely of selected-response items.[21] Although this can solve the problems associated with constructed-response anchors it does so at the risk of violating the requirement that the anchor fully represent the construct measured by the entire test. This can, as a consequence, bias the resulting equating (Kim & Kolen, 2006). The success of this approach depends most importantly on the multiple-choice and constructed-response components of the test being highly correlated (Kim et al., 2010).

Equivalent Groups Equating of Performance Assessments

Perhaps the best way of addressing the challenges performance assessments present to common-item equating methods is to avoid those methods altogether. The equivalent groups design is an attractive alternative because it sidesteps concerns regarding the construct representativeness of an anchor. As illustrated in Figure 3-1, test forms A and B (which need not share items) are both administered to randomly equivalent groups. This is most often done by administering both forms on the same occasion with the test takers randomly assigned to each. Because the test taker groups are equivalent, any differences in performance across them can be attributed solely to differences in the test forms themselves (Kolen & Brennan, 2004).

A typical application of the equivalent-groups design would involve administering both an old and a new form to randomly sampled test takers on a given occasion. It is not necessary that the two forms be administered to equally sized samples (although the accuracy of the equating is driven by the size of the smaller sample). For example, security considerations might suggest that the old form, which has been used previously, be administered to a much smaller sample than the new form. The new form on Occasion 1, then, becomes the old form on Occasion 2 in order to bring yet another new form onto scale. A twist on this design would reverse the large and small samples and designate the old form as the principal form on each occasion. This would mean that the bulk of the test takers tested on each occasion were taking a form that was previously equated and so could be scored without the delay required to perform the test equating analyses.

[21] Tests composed solely of constructed-response items may also be equated through multiple-choice anchors that are external to the test itself. For example, a constructed-response writing test might be anchored by a separately administered and independently scored multiple-choice writing test.

Although the equivalent-groups design solves most of the problems associated with selecting and administering anchor tasks under performance assessment, it still requires that the tasks administered on occasion one (which now compose the full test form) perform identically on occasion two. Concerns regarding rater consistency and test-taker-by-task-by-occasion interactions therefore remain valid. Moreover, this approach relies only on judgment that the two forms measure the same thing (as the same test taker does not take both forms for empirical comparison). Consequently it would be desirable to check that scores from the two forms have the same correlations with external measures, such as scores on assessments of other subjects or prior-year scores from the same subject.

Single-Group Equating of Performance Assessments

The single-group, nearly equivalent test (SiGNET) design also holds promise for equating even tests comprising a small number of constructed-response tasks. This design was developed for use under very challenging conditions: testing programs that test relatively few test takers but do so with frequent or even continuous administrations (Grant, 2011). Low test-taker volume and frequent administration is a poor fit to traditional equivalent-groups designs because scores could not be provided immediately to those test takers presented the new test form. They would have to wait until sufficient data were collected to conduct the equating. The SiGNET design solves that problem by replacing the old test form with the new in stages rather than at once, equating the two gradually as data are collected. This is illustrated by Figure 3-2.[22]

Form	Tasks				
A	Task 1	Task 2	Task 3	Task 4	Task 5*
B	Task 6*	Task 2	Task 3	Task 4	Task 5
C	Task 6	Task 7*	Task 3	Task 4	Task 5
D	Task 6	Task 7	Task 8*	Task 4	Task 5
E	Task 6	Task 7	Task 8	Task 9*	Task 5

*Scores on this task are not included in the examinees' total scores for this form.

Figure 3-2. The SiGNET design.

The initial test form (A) consists of four scored tasks (1-4) and one new, unscored task (5). It is assumed that test scores summed across the four scored tasks either constitute the base or reported score scale or have been previously linked to that base. Once sufficient data have been collected on Form A, an equating of a particular sort can be conducted. This is to link scores aggregated across Tasks 1 through 4 with those aggregated across Tasks 2 through 5. Because all test takers have responded to all five tasks, this is effectively a single-group equating. Because the single group is a robust equating design even modest sample sizes can produce stable results (Kolen & Brennan, 2004; Livingston, 2014). The outcome of this equating allows scores on a new form consisting of Tasks 2 through 5 to be linked to the base scale represented by Tasks 1 through 4. This new form is designated as B and can then be administered, now accompanied by a new unscored Task 6. Because Form B has been equated when it was administered as part of Form A, scores can be reported without delay, a traditional advantage of pre-equating designs. Note that the equating of Forms A and B can be conducted by any of the standard classical or IRT-based methodologies. This process continues through Forms C and D, successively replacing older

[22] The new task is always shown in the last position in this diagram for the sake of simplicity. In practice, task position would be kept more stable across time to better minimize form differences. For example, Task 5—which replaces Task 1—would be pretested in the second position, moving up to the first in Form B when Task 1 is removed.

tasks with newer. By the time Form E appears, the initial Form A has been entirely replaced and the cycle begins anew.

The SiGNET design can be seen as a variant of a design under which new tasks are pretested and IRT calibrated alongside an existing operational form. Once sufficient tasks have been calibrated, a new form can be assembled and administered, replacing the old form at once rather than successively over time. The SiGNET design is based on very strong single-group equating, making it feasible with sample sizes of only a few hundred test takers per form (Grant, 2011). Changing the operational test form continuously (if incrementally) over time rather than leaving the initial form in the field intact until it can be replaced entirely may also improve test security under some circumstances.

Task and Test Scoring

Several actions can be taken to control the effects of rater inconsistency during the scoring process. Most of these actions attempt to manage the performance of human raters through extensive training, double scoring, adjudication of discrepant scores and use of comparability or trend scoring. Assigning two or more raters to score some or all of each test taker's responses both increases the reliability of scoring and facilitates use of a wide range of rater quality-control measures. Having responses scored by multiple raters allows agreement levels across raters to be tracked. Coupling this with comparisons of each rater's distribution of scores relative to other raters can provide a detailed view of rater performance.

More recently it has become possible to compare raters not just with one another but also with the output of automated scoring systems. When properly implemented, these systems can provide a stable benchmark over time, partially mitigating rater by task by occasion interactions and rater drift on the rating scale.

Statistical modeling of rater performance can also prove effective. Various item response models that include parameters for test taker ability, task difficulty and rater severity have been proposed and applied (e.g., de Ayala, 2014). These models can be used both to evaluate rater performance and to adjust test-taker scores statistically for differences in task difficulty and rater severity.

Computerized Scoring of Performance Assessments

It has long been hoped that computerized scoring of performance assessments can make their use both more practical and more psychometrically sound (Page, 1966). Freed from the time and expense inherent in organizing and managing large teams of human raters, performance assessments could conceivably be scored as quickly and cheaply as multiple-choice tests. And if computerized scoring were consistent across time and impartial across test takers, it could eliminate rater variability as a source of score incomparability. Although these hopes have not yet been fully realized, the promise remains (Shermis, Burstein, Brew, Higgins, & Zechner, in press).

Computerized scoring has proven largely successful in several limited domains (Williamson, Mislevy, & Bejar, 2006). The first is the scoring of writing samples, for which a substantial research base has been established. Computer-generated scores are generally found to agree quite strongly with human ratings of writing quality (e.g., grammar, usage, mechanics, organization, some elements of style). However, computerized scoring systems struggle or fail entirely in evaluation of writing content. For example, systems are generally incapable of evaluating whether a thesis was logically argued or a position persuasively defended or even supported by relevant examples (Shermis & Burstein, 2013).

Computerized scoring has also been successful in evaluating performance on specific but complex tasks. Notable examples of these tasks include the design and placement of architectural structures and the management of medical patient scenarios (Bejar & Braun, 1999; Clauser, Margolis, Clyman, & Ross, 1997). In both of these examples, the goal of computerized scoring is again to emulate the performance of human raters.

Because computerized scoring may not yet be sensitive to all aspects of test taker performance there is the real risk of either narrowing the construct being measured or even introducing construct-irrelevant variance. For example, most essay-scoring algorithms are predisposed to reward long essays cluttered with low-frequency vocabulary even if neither of these characteristics improves the quality or readability of the writing (Bridgeman, Trapani, & Attali, 2014). Worse, their focus on particular features of writing can leave scoring algorithms vulnerable to gaming by test takers who intentionally exaggerate these features in their writing (Powers, Burstein, Chodorow, Fowles, & Kukich, 2001; Perelman, 2012).

Despite current shortcomings, computerized scoring systems can be used effectively to monitor and stabilize the performance of human raters if not replace them outright. One approach is to compare the scores of the computerized system with those of a single human rater. If a sufficiently large discrepancy is observed, a second human rater would be brought in to adjudicate. Because computerized scoring can be made stable over time it can serve much the same purpose as trend scoring where responses scored by past human raters are interspersed with new responses to determine whether the current rater cohort is consistent with scoring in the past (Trapani, Bridgeman, & Breyer, 2014).

Matrix Sampling of Tasks

For the purpose of estimating performance at an aggregate level, such as school performance, matrix sampling of tasks is an effective way to reduce the number of tasks performed by individual test takers to a manageable level and still maintain reliability of scores at the school level. For example, different random samples of test takers within a school may be administered different collections of tasks. In a comparison of the generalizability of school-level scores using a crossed design in which all students completed the same tasks and a matrix-sampling design in which different groups of students within each school completed different test forms, Gao, Shavelson, and Baxter (1994) found that matrix sampling reduced the number of tasks needed per student by more than half.

Conclusions

Three fundamental challenges to the comparability of performance assessment scores were identified and described. First, the nature of constructed-response tasks both complicates the assembly of parallel test forms and makes it less likely that these forms will perform comparably across test takers and test occasions. Second, the rating process required to score constructed-response tasks risks introducing sources of irrelevant score variability. Finally, variability in form performance across occasions also threatens score comparability. A number of mitigation strategies was then identified for coping with each of these challenges:

1. Extend the test length. The simplest and often most effective way of improving score comparability (and increasing score reliability) is to increase the length of a test. The substantive and statistical characteristics of performance assessments that limit score comparability become progressively less problematic as test length increases. There are, of course practical and policy consequences of increased test length. The question is therefore whether a test's sponsors and stakeholders value (or require) score comparability more than shorter testing time.

2. *Augment performance assessments with selected-response items*. Performance tasks are often integrated into a mixed-format assessment that also includes selected-response or similarly machine-scored items (e.g., technology enhanced). Such items are perfectly capable of measuring certain aspects of many constructs. Because selected-response items generally require far less administration time than performance tasks, their use can effectively lengthen the test (with the attendant benefits described above) without substantially increasing testing time. Selected-response items are also a practical, if not entirely satisfactory, choice for anchoring a common-item equating.

3. *Rigidly standardize tasks and test forms*. Form comparability certainly benefits from detailed task models and assembly specifications. Even if these efforts are unlikely to make form performance so stable as to render statistical adjustments of scores superfluous, they are likely to strengthen any equating methods that may be applied.

4. *Accept that scores are not fully comparable and limit inferences accordingly*. Lower levels of score comparability may simply need to be accepted as the price for measuring otherwise inaccessible constructs or for measuring in more direct ways. It may also be the case that the circumstances under which a test is used do not require comparability across forms. The example was given of a test that is administered just once a year to a cohort of test takers seeking admission to an academic or professional program. Provided any decisions made required comparison only among test takers tested in the same cohort, high levels of score comparability across test forms and administration occasions is not required.

5. Less than fully comparable scores may also be perfectly acceptable for use in lower stakes circumstances. For example, a performance assessment used for formative purposes may provide very adequate instructional support without high levels of comparability.

6. *Report only group-level scores.* Group-level inferences can remain valid even when the scores achieved by individual test takers are not comparable. Indeed, group-level inferences may be strengthened by sacrificing comparability of individual scores. Matrix-sample based testing programs are the best example of this. These programs essentially develop very long test forms in order to ensure that a content domain is thoroughly and properly sampled. Because these forms are too long to be reasonably administered intact, they are instead divided into multiple blocks and distributed across test takers. Because test takers take what may be nonequivalent item blocks, comparability at the test taker level is limited. However, aggregating results up to the group level can produce very stable results across time and changing samples of items, largely due to the very long form lengths. This design is the basis of a number of large-scale educational surveys, including NAEP, PISA, and PIRLS (Jones & Olkin, 2004).

References

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12,* 238–257.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London, England: Arnold.

Baxter, G. P., & Shavelson, R. J. (1994.) Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research*, *21*, 279–298.

Bejar, I. I., & Braun, H. I. (1999). *Architectural simulations: From research to implementation: Final report to the National Council of Architectural Registration Boards* (Research Memorandum No. RM-99-02). Princeton, NJ: Educational Testing Service.

Bridgeman, B, Trapani, C., & Attali, Y. (2014). Comparison of human and machine scoring of essays: differences by gender, ethnicity, and country. In C. Wendler & B Bridgeman (Eds.), *The research foundation for the GRE® revised General Test: A compendium of studies* (pp. 4.8.1–4.8.3). Princeton, NJ: Educational Testing Service.

Clauser, B. E., Margolis, M. J., Clyman, S. G., & Ross, L. P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational Measurement, 34*, 141–161. doi: 10.1111/j.1745-3984.1997.tb00511.x

Cronbach, L. J. (1990). *Essentials of psychological testing.* New York, NY: Harper & Row.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*, 373–399.

de Ayala, R. J. (2014). The IRT tradition and its application. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (vol. 1). London, England: Oxford University Press.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70,* 287–322.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11*, 195–208.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7,* 323–342.

Grant, M. C. (2011). *The single group with nearly equivalent tests (SiGNET) design for equating very small volume multiple-choice tests* (Research Report No. RR-11-31). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.

Jones, L. V., & Olkin, I. (Eds.). (2004). *The nation's report card: Evolution and perspectives*. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5–17.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*, 357–381.

Kim, S., Walker, M. E., & McHale, F. (2010), Investigating the effectiveness of equating designs for constructed-response tests in large-scale assessments. *Journal of Educational Measurement, 47*, 186–201.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., … Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11,* 121–137.

Kolen, M. J., & Brennan, R.L. (2004). *Test equating, scaling and linking.* New York, NY: Springer-Verlag.

Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment instrument*. Educational Measurement: Issues and Practice, 12*(2), 16–23.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33,* 71–92.

Lane, S., & Stone, C. A. (2006). Performance assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.; pp. 387–432). Westport, CT: Praeger.

Liu, X., & Li, L. (2014). Assessment training effects on student assessment skills and task performance in a technology-facilitated peer assessment. *Assessment & Evaluation in Higher Education, 39*, 275–292.

Livingston, S. A. (2014). *Equating test scores (without IRT)* (2nd ed.). Princeton, NJ: Educational Testing Service.

Maryland State Department of Education. (1996, March). *1996 MSPAP public release task: Choice in reading and writing scoring guide.* Baltimore, MD: Author.

McBee, M. M., & Barnes, L. L. B. (1998). The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education, 11,* 179–194.

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2014). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice.* Advance online publication. doi:10.1111/emip.12061

Moss, P. A., Cole, N. S., & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at grades 4, 7, and 10. *Journal of Educational Measurement, 19,* 37–47.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238–243.

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 121–132). Anderson, SC: Parlor Press.

Powers, D. E., Burstein, J., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2001). *Stumping e-rater: Challenging the validity of automated essay scoring* (GRE Report, No. 98-08bP). Princeton, NJ: Educational Testing Service.

Richter Lagha, R., Boscardin, C. K., May, W., & Fung. C. C. (2012). A comparison of two standard-setting approaches in high-stakes clinical performance assessment using generalizability theory. *Academic Medicine, 87,* 1077–1082.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*, 41–53.

Shavelson, R. J. (2010). *Measuring college learning responsibly: Accountability in a new era.* Stanford, CA: Stanford University Press.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215–232.

Shavelson, R. J., Mayberry, P. W., Li, W., & Webb, N. M. (1990). Generalizability of job performance measurements: Marine Corps Riflemen. *Military Psychology, 2*, 129–144.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36,* 61–71.

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge.

Shermis, M. D., Burstein, J., Brew, C., Higgins, D., & Zechner, K. (in press). Recent innovations in machine scoring. In S. Lane, T. Haladyna, & M. Raymond (Eds.), *Handbook of test development* (2nd ed). New York, NY: Routledge.

Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice*, *16*(3), 16–24.

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffery, D. F., Robyn, A., Shavelson, R. J., & Haertel, E. (2000). The effects of content, format, and inquiry level on science performance assessment scores. *Applied Measurement in Education, 13*, 139–160.

Steedle, J. T., & Elliot, S. (2012, April). *The efficacy of automated essay scoring for evaluating student responses to complex critical thinking performance tasks*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement*, *37*, 329–346.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68,* 249–276.

Trapani, C., Bridgeman, B., & Breyer, J. (2014). Using automated scoring as a trend score: The implications of score separation over time**.** In C. Wendler & B. Bridgeman (Eds.), *The research foundation for the GRE revised General Test: A compendium of studies* (pp. 2.4.1–2.4.4). Princeton, NJ: Educational Testing Service**.**

van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika, 65*, 437–456.

van Zundert, M., Sluijsmans, D., & van Merrienboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20,* 270–279.

Webb, N. M., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education, 13,* 277–301.

Webb, N. M., Shavelson, R. J., Kim, K. S., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinists mates. *Military Psychology, 1*, 91–110.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*, 2–13.

IV. Performance Assessment: Comparability of Groupwork Scores[23]

Introduction 58
What Is Groupwork? 58
Why Might We Incorporate Groupwork Into Performance Assessments? 59
Preparing Students for Individual Performance 59
Measuring Groupwork Processes and Outcomes 59
Linking Assessments to Instruction 60
Example Assessments With Groupwork 60
Concerns About Construct Interpretations: The Effect of Group Processes 61
Task-Related Interactions 62
Lack of Involvement 62
Uncoordinated Group Communication 62
Social-Emotional Processes 63
Division of Labor 63
Sources (Facets) of Measurement Error Specific to Groupwork 63
Variation Due to Group Composition 64
Variation Due to Role Assignment in Groupwork 66
Variation Due to Type of Task 66
Variation Due to Task 67
Variation Due to Occasion 67
Variation Due to Type of Rater 67
Variation Due to Type of Rating or Rating Scale 68
Agreement Among Raters 69
Automatic Coding and Scoring of Group Processes 69
Variation Due to Random Coding Errors and Data Entry Mistakes 70
Implications for Validity and Reliability 71
Additional Issues Introduced by the Use of Groupwork in Assessments 72
Relationships Between Process and Productivity/Performance 72
Multiple Levels of Objects of Measurement: Individual and Group 73
Conclusions 74
References 75

[23] Noreen Webb was lead author for Chapter IV.

Introduction

In defining what students should understand and be able to do to be prepared for college and careers in the 21st century, current curriculum standards explicitly address the importance of communicating and collaborating with others. The standards for speaking and listening in the Common Core State Standards for English language arts & literacy, history/social studies, science, and technical subjects, for example, call for students to have opportunities to participate in conversations with others in large and small groups to learn how to work together and to develop useful oral communication and interpersonal skills (National Governors Association Center for Best Practices, & Council of Chief State School Officers, 2010a). These skills include being able to express ideas and listen carefully to others' ideas. Similarly, the Common Core State Standards for mathematics stress communicative competence and specifically highlight the need for students to be able to justify their ideas and communicate them to others and to respond to others' arguments (National Governors Association Center for Best Practices, & Council of Chief State School Officers, 2010b).

Paralleling the use of group settings in the classroom for communication and collaboration, measuring students' ability to communicate and collaborate with others requires groupwork on assessments. Performance assessment, then, can extend beyond the measurement of purely individual performance and involve individuals working in groups. Although performance assessments in K-12 education have rarely included groupwork, performance assessments of job readiness and job performance in the workplace routinely incorporate groupwork for the purpose of measuring individuals' ability to work in teams, as is discussed below.

In addition to the usual challenges of measuring individuals' performance on complex tasks, groups working on complex tasks present additional measurement challenges. Here we enumerate and describe these challenges. We conclude with suggestions for addressing the challenges and by identifying areas for future research.

What Is Groupwork?

The preceding chapter addressed individuals' performance ostensibly working alone. In this chapter, we consider individuals working together in small groups of size two or more to achieve a common goal, whether it is solving a problem, completing a task, or producing a product. The object of the measurement may be the individual or the group. Moreover, the measurement may focus on processes occurring during the group's work, the product generated by the group, or performance exhibited during or after groupwork. Specifically, interest may lie in one or more of the following: (a) an individual's group-process competencies; (b) the quality of an individual's product, performance, or problem solution after working in a group; (c) the nature and variability of group processes on particular task(s); or (d) the quality of a group's solution to a problem, performance on a task, or product generated.

We conceive of groupwork quite broadly. Interactions among members of a group may occur face-to-face or virtually and may occur synchronously or asynchronously. Group members may or may not play specific roles or be assigned specific activities, and groups may or may not follow scripts or instructions about how they should interact. Moreover, not all members of the group need be human; some may be computer agents playing the role of fellow examinee. We also consider situations in which humans or computer confederates play the role of a nonpeer (e.g., trained interviewers or computer agents playing the role of an expert or native-speaking conversational partner in tests of oral language proficiency; trained actors playing the role of patient in tests of clinical skills in medicine).

Why Might We Incorporate Groupwork Into Performance Assessments?

Performance assessments might include groupwork for a variety of reasons. We consider three main purposes: (a) to prepare students for individual performance, (b) to measure processes and outcomes of groupwork, and (c) to link assessments to instruction.

Preparing Students for Individual Performance

Performance assessments might include groupwork as preparation for subsequent individual performance. A groupwork phase that precedes a purely individual phase may be used to help equalize intellectual resources for students who have had less opportunity than other students to learn the material. That is, the opportunity to share knowledge and understanding during groupwork practice may help level the playing field (Baron, 1994; Neuberger, 1993). Including preparatory groupwork prior to individual work may also make it possible to measure how well students can learn from working with others, which is consistent with a perspective of student competence that sees learning as constructed in collaboration with others (e.g., Vygotsky, 1978). Providing opportunities for groupwork and then assessing students after groupwork practice, then, may be seen as a way both to increase fairness and to measure how well students can learn from collaborative experiences. For many reasons, however, groupwork practice may not necessarily achieve these goals of increasing fairness and providing learning opportunities, as will be described in later sections.

Groupwork was used as preparation for individual work in past assessments. For example, in the Maryland School Performance Assessment Program (MSPAP, Maryland State Department of Education, 1994), science assessments had students work in 4- or 5-person groups to perform an investigation and then individually answer test questions that were based on a set of hypothetical data for the same experiment; language arts assessments had students discuss background information in groups (e.g., jobs performed by community workers) and then individually write essays using and extending what they had discussed in their group. The Smarter Balanced Assessment Consortium currently uses preparatory groupwork in formative assessment tasks in which students discuss concepts and ideas in small groups and then write individual responses to questions based on those concepts. In assessments with preparatory groupwork, students' discussion and work generated during the group phase is not scored.

Measuring Groupwork Processes and Outcomes

Another reason for including groupwork in performance assessments is to measure students' abilities to collaborate with others and to accomplish tasks in a group. In the workforce, organizations increasingly rely on groups or teams to carry out many types of tasks (e.g., production, service, management, project, action and performing, advisory, negotiation, problem solving) in a wide variety of settings (e.g., military, government, civilian; Wildman et al., 2012). Large or complex tasks require teams to complete them successfully (e.g., designing new products); other tasks, by definition, involve multiple participants (negotiating agreements). Employers view the ability to collaborate with others to accomplish tasks as a core 21st century competency that is more important than even English language ability or subject matter knowledge for both landing and keeping a job (National Research Council, 2011). Teamwork skills seen as essential for successful group performance and high-quality group productivity include, for example, adaptability (recognizing problems and responding appropriately), communication (exchanging clear and accurate information), coordination (organizing team activities to complete a task on time), decision-making (using available information to make decisions), cooperative interaction with other team members (interpersonal), and leadership (providing structure and direction for the team, Chung, O'Neil, & Herl, 1999; Salas, Sims, & Burke, 2005; SCANS, 1999).

These teamwork skills have much in common with the communication and collaboration skills deemed to be essential features of the curriculum standards for K-12 education, as described above.

Incorporating groupwork into assessments provides a direct way of measuring both collaborative skills and the outcome of the group's efforts (the group's product or performance). As described below, groupwork can generate scores for both processes and outcomes and for both individuals and groups. Individual members of the group can be scored on their collaboration skills and on their contributions to the group's product. And the group as a whole can receive scores for its functioning (group process) and its outcome (product or performance).

Linking Assessments to Instruction

Finally, performance assessments may include groupwork to link assessments to instruction. First, including groupwork on assessments constitutes a signaling function for classroom instruction (Linn, 1993). Decades of research in instructional settings show the power of collaborative groupwork for learning and other outcomes such as the development of prosocial attitudes (see Webb & Palincsar, 1996, for a review of research). Incorporating groupwork into assessments can serve as a policy lever to influence classroom instructional practices, especially if the assessments simulate a beneficial learning environment that affords high-quality collaboration.

Second, incorporating groupwork into assessments may increase their fidelity to instructional activities (and thus increase their instructional sensitivity or instructional validity, Popham, 2007). Instructional activities incorporate groupwork to provide opportunities for collaboration, to afford opportunities to tackle large and complex tasks that individual students cannot easily complete, or both. Groupwork activities in the classroom have the potential to confer multiple benefits. For example, including partner work on science laboratory tasks in the classroom (e.g., investigating why an ice cube sitting on an aluminum block melts faster than an ice cube on a plastic block) ideally provides opportunities for students to engage in collaborative scientific argumentation that fosters scientific literacy (e.g., discussions about gathering and making sense of data, generating and testing hypotheses, justifying explanations, critiquing viewpoints, Sampson & Clark, 2011). Group preparation of multifaceted research reports on complex topics, such as the phototropic behavior of plants, is another example of groupwork with the potential for the exchange of ideas and development of new knowledge and understanding (Cohen & Cohen, 1991). Assessment tasks that provide collaborative opportunities may represent such instructional activities better than purely individual tasks do.

Example Assessments With Groupwork

A number of large-scale assessments have incorporated groupwork (or will do so) to serve one or more of the purposes described above. For example, in the 1990s, several state assessments incorporated collaborative groupwork alongside individual assessments in response to recommendations of state and national assessment standards. One was the three-part science task in the Connecticut Common Core of Learning Alternative Assessment Program, in which students first individually provided information about their science knowledge; then worked in 3- or 4-person teams to design, carry out, interpret, and summarize an experiment; and finally individually reflected about the group activity, independently analyzed and critiqued another group's report, and applied knowledge gained during the groupwork phase (Baron, 1992, 1994). Students' performance was scored on all three parts. Other assessments with collaborative groupwork included the Connecticut Academic Performance Test (CAPT, Connecticut State Board of Education, 1996; Wise & Behuniak, 1993), the Maryland School Performance Assessment Program (MSPAP, Maryland

State Department of Education, 1994), the California Learning Assessment System (CLAS, Saner, McCaffrey, Stecher, Klein, & Bell, 1994), and the 1994 Kansas Science Assessment (Pomplun, 1996).

Currently, the Smarter Balanced Assessment Consortium uses groupwork in formative assessment tasks. For example, in a sample Grade 11 Performance Task on Thermometer Crickets (Smarter Balanced Assessment Consortium, 2012), students work in small groups during classroom instruction to build background knowledge (e.g., why crickets chirp primarily at night), engage in a class discussion (e.g., interpretation of data about cricket chirping in different conditions), and then complete tasks individually (e.g., organizing and analyzing data about the relationship between temperature and crickets' chirping rates). The scoring rubric focuses on student performance on the individual task that follows the class discussion (e.g., plotting the data points, modeling and interpreting the relationship). Groupwork appears internationally in high-stakes assessments as well, such as the Singapore A-levels required for university admission (National Research Council, 2011).

In the near future (planned for 2015), another international assessment—the Programme for International Student Assessment's (PISA) test of individuals' collaborative problem-solving skills—will have examinees interact with one or more computer simulated collaborators to solve problems such as finding the optimal conditions for fish living in an aquarium or producing an award-winning logo for a sporting event (Organisation for Economic Co-operation and Development, 2013). Examinees will be scored on their behavior when interacting with the computer agent (e.g., communicating about the actions to be performed to complete the task) and on their responses to multiple-choice and constructed-response probes placed within the unit (e.g., write an email explaining whether there is group consensus on what to do next). The scoring will focus on three competencies: establishing and maintaining shared understanding, taking appropriate action to solve the problem, and establishing and maintaining team organization. Specific skills within these dimensions include, for example, coordination, explanation, filling roles, argumentation, and mutual regulation. Along similar lines, the National Center for Education Statistics within the U.S. Department of Education is also planning how to assess collaborative problem solving in the National Assessment of Educational Progress (2014 NAEP Innovations Symposium: Collaborative Problem Solving held in Washington, DC, September 29, 2014).

Concerns About Construct Interpretations: The Effect of Group Processes

Groupwork introduces complexities not found in purely individual assessments. Regardless of the reason(s) for incorporating groupwork into an assessment, the mere presence of groupwork renders invalid an interpretation of scores as reflecting unassisted individual competence. For example, in the 90-minute Connecticut Academic Performance Test in language arts, insertion of a brief 10-minute group discussion partway through the test improved students' understanding of the story and their scores on the test (Fall, Webb, & Chudowsky, 2000). Similarly, designing and carrying out science experiments and investigations in pairs on the California Learning Assessment System helped students develop new ideas, knowledge, and understanding (Saner et al., 1994). Students' scores on the tests, then, reflected a combination of their own competence and what they learned or gained from the groupwork experience.

Further complicating the interpretation of scores from assessments with groupwork, some groups function more effectively than others for reasons that may or may not align well with the construct(s) of interest. Several ways in which group functioning may differ, with consequences for assessment scores, include:

- Task-related interaction with others
- Lack of involvement
- Uncoordinated group communication
- Social-emotional processes
- Division of labor

Whether these processes are beneficial or detrimental for performance scores will depend on the target construct.

Task-Related Interactions

Group members can interact with each other around the task in a great many ways, such as sharing information and ideas; building on each other's ideas to solve a problem, constructing new knowledge, or completing a task; engaging in conflicts and resolving disagreements; and seeking, giving, and receiving help (Webb & Palincsar, 1996). Giving and receiving help, for example, can promote learning, and thus improve individuals' scores after groupwork practice, by encouraging students to rehearse information, reorganize and clarify material in their own minds, recognize misconceptions and gaps in understanding, strengthen connections between new and previously learned information, and develop new perspectives. Engaging in helping behavior may also lead raters to score examinees highly on communications skills. In terms of group productivity, however, spending time to ensure that everyone understands the material may slow the group down and prevent it from solving the problem or completing the task. In that case, suppressing the participation of less capable members or students who are experiencing difficulty may help the group improve its performance score.

Lack of Involvement

Members of a group may not participate if they are discouraged, unmotivated, unrecognized, intellectually snobbish, intentionally passive, or involved in something else (Mulryan, 1992). Consider, for example, social loafing, or diffusion of responsibility, which arises when one or more group members sit back and let others do the work (Karau & Williams, 1993; Slavin, 1990). Individuals may go along for a free ride if they believe that their efforts cannot or will not be identified or are dispensable (Kerr & Bruun, 1983; Levine & Moreland, 1990). Uninvolved students will likely receive low individual scores on their subject-matter communication skills and contributions to teamwork and possibly also individual scores after groupwork practice, especially if they lacked relevant knowledge or skills coming into the assessment (Webb, 1993). How social loafing might affect *group* scores depends on group members' capabilities and task attributes. On the one hand, social loafing may be detrimental for group productivity if the social loafers have necessary skills for the group to accomplish the task (which may be especially relevant for nonroutine tasks that do not have well-specified procedures, Cohen & Cohen, 1991), or if social loafing becomes contagious (Salomon & Globerson, 1989). On the other hand, groups may function better and complete tasks more effectively if some students keep quiet, especially if they do not have new or productive ideas to contribute.

Uncoordinated Group Communication

Instructional research shows that opportunities for groups to benefit from information sharing may be lost when group members do not coordinate their communication. In uncoordinated conversations, students advocate and repeat their own positions and ideas, ignore others' suggestions, reject others' proposals without elaboration or justification, and interrupt others or talk over them (Barron, 2000). In highly coordinated groups, in contrast, members acknowledge and elaborate upon each other's ideas. Although lack of coordination of group members' efforts on assessments with groupwork can impede group functioning

and reduce the quality of the group's product (and thus their group productivity score), students who actively promote their own ideas (even if they do not engage with others) may nonetheless receive high individual communication scores.

Social-Emotional Processes

Negative social-emotional processes, such as being rude, hostile, unresponsive, and domineering, can impede group functioning in multiple ways, such as causing group members to withhold correct information from each other and to reject viable suggestions posed by others (Chiu & Khoo, 2003). While such processes can negatively affect group productivity and reduce opportunities for group members to benefit from groupwork practice (Webb & Mastergeorge, 2003), these processes may not be detrimental for individuals' communication scores (e.g., dominant students being marked high for their frequent contributions). Positive social-emotional processes such as cooperativeness, cohesiveness, team spirit, and liking of team members may improve group productivity and performance unless the good feelings arise out of suppression of disagreements, which can lead to reduced group productivity and opportunities to benefit from groupwork practice (Webb & Palincsar, 1996).

Division of Labor

Division of labor, that is, dividing the task into parts and assigning different group members responsibility for completing different parts, may be a productive, efficient, or even necessary, strategy for accomplishing group tasks (Salas, Burke, & Cannon-Bowers, 2000) and may, consequently, increase group performance or group productivity scores. However, if this strategy curtails interaction among group members, it may produce underestimates of individuals' scores on, for example, their ability to collaborate with others, communicate about the subject matter, and apply or synthesize knowledge gained during groupwork practice.

In summary, the nature of group processes that arise in a particular groupwork session may greatly impact scores of groups and/or their members. Some influences may be construct-relevant (a group with highly coordinated communication receives a high score on teamwork skills), while other influences may be construct-irrelevant (a student receives a low communication score because the group divided up the labor to accomplish the task and spent little time discussing it). Next we consider influences on processes and performance in the groupwork setting that may cause scores to vary and, consequently, affect validity and/or reliability of score interpretations.

Sources (Facets) of Measurement Error Specific to Groupwork

The previous chapter enumerated a number of important sources of unwanted variation in individual-level assessment scores, such as variability due to the sampling of tasks, occasions, raters, rater types, and type of rating or rating scale. All of these sources (facets) of measurement error figure prominently in assessments with groupwork as well. In addition, the groupwork setting introduces new sources of score variability that do not pertain to individual assessments, such as the composition of the group and the roles played by group members. Here, we address sources of measurement error that are unique to groupwork, as well as ways in which previously mentioned sources of error may operate in new or different ways in groupwork settings. We draw from research on groupwork in education and (where relevant) other fields such as organizational psychology. Furthermore, we consider multiple assessment contexts in which sources of measurement error of scores from groupwork have been studied, including schools, industry, the military, and medicine.

Variation Due to Group Composition

The composition of the group can vary along a great many dimensions including group-member knowledge and skill level, gender, personality, motivation, acquaintanceship, status, popularity, attractiveness, ethnic background, race, and other demographic characteristics. The large instructional literature on cooperative or collaborative learning in the classroom shows a marked influence of group composition on many outcomes, including group processes, group performance, and student learning. Similarly, research in organizations (e.g., industry, military) shows that team composition on, for example, cognitive and psychomotor abilities, organizing skills, cooperativeness, team orientation) greatly influences team functioning and success (National Research Council, 2013; Mathieu, Tannenbaum, Donsbach, & Alliger, 2014). Emerging evidence shows that group composition, especially the homogeneity of the group in terms of, for example, student achievement level or perceptions about the task or teamwork skills, matters in assessment situations, too. Some studies, for example, have found homogeneous groups to produce higher scores than heterogeneous groups on collaborativeness, resolution of cognitive conflicts, and communication when engaging in complex mathematics tasks (Fuchs, Fuchs, Hamlett, & Karns, 1998) and aircraft simulation tasks (Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000).

Of particular interest is the *combined* influence of group composition and group members' own characteristics on group processes and outcomes, such as average-ability students, low-status students, and girls being less active and learning less in heterogeneous than in homogeneous classroom groups (Webb & Palincsar, 1996). This combined influence appears in assessments studies as well, such as high-ability students showing higher scores in homogeneous groups and in high-functioning heterogeneous groups than in poorly functioning heterogeneous groups (e.g., group members failing to answer each other's questions, failing to explain their answers, insulting others, Webb, Nemer, & Zuniga, 2002).

Growing concern about the combined roles of test-taker characteristics and group composition appears in recent research on testing the language capability of students learning English as a second language. Increasingly, group oral tests are being introduced into low-stakes and high-stakes tests in order to assess communication ability in a more naturalistic setting than, say, an oral interview with an examiner. Characteristics such as gender, personality (especially introversion-extraversion), language proficiency, and acquaintanceship (friends, acquaintances, and strangers) have different effects on communication depending on how test takers are paired (e.g., Ockey, Koyama, & Setoguchi, 2013), as well as on the size of the group (e.g., introverted students participate more actively in smaller groups than in larger groups, Nakatsuhara, 2011).

Recent evidence suggests that behavior of, and scores assigned to, the same individual may change from one group composition to another. One example is a study conducted in a managerial assessment center, commonly used in the business community to gauge prospective employees' job skills such as communication, consideration/awareness of others, drive, influencing others, organization and planning, problem solving, leadership, learning from others, fostering relationships, and managing conflict (Collins & Hartog, 2011). Assessment center exercises include, for example, role plays in which an examinee presents a business plan or carries out a coaching conversation with a trained role player who responds in prescribed ways to the test taker's actions. Hoffman and Meade (2012) observed managers in an executive MBA program in two such role play exercises. In one, test takers interacted with a high-performing but interpersonally abrasive role player; in the other, they interacted with an average performing (and presumably nonabrasive) role player. Ratings of group process dimensions (e.g., oral communication, sensitivity, leadership, confrontation) for the two role play exercises did not correlate highly, showing that scores may not be generalizable from one group composition to another, and that test

takers may need to be observed in a large number of group compositions for generalizable results.

Given the difficulty of ensuring that group compositions are similar across groups, considerable interest lies in controlling variation in group composition by standardizing attributes and behavior of group members. Hoffman and Meade's (2012) study points to one method of standardization: using scripted confederates to play the role of group partners. Scripted group members have been used in a variety of assessment situations, such as live actors playing the role of patients in medical performance tasks involving history taking, physical examination, and patient education (e.g., Richter Lagha, Boscardin, May, & Fung, 2012), confederates posing as team members (co-pilots) and following a script that presents prescribed conflict situations for test takers to resolve in flight simulation tasks, such as making blatant errors and remaining indecisive regarding critical decisions (Smith-Jentsch, Salas, & Baker, 1996), and trained interviewers engaging in structured conversations with test takers in tests of oral proficiency (e.g., the American Council on the Teaching of Foreign Languages Oral Proficiency Interview[24]).

Unless the behavior of confederates is highly controlled, however, their attributes and behaviors can introduce error into the measurement. Lazaraton (1996), for example, documented multiple ways in which interviewers in oral proficiency assessments might influence conversations (and, hence, test takers' language scores), such as interviewers completing test takers' sentences or thoughts, echoing or correcting test-taker responses, repeating questions using slowed or over-articulated speech, or rephrasing questions. On the other hand, tightly scripting confederates to eliminate such influences may produce unnatural conversational discourse (Malone, 2003).

As an alternative to using human confederates, test takers might interact with computer conversational agents (simulated group members) that are programmed to respond to test taker behavior in certain ways. The use of computer agents is at the heart of the 2015 PISA effort to measure collaborative problem solving competencies (Organisation for Economic Co-operation and Development, 2013). Conversational agents will represent peers with a range of skills and abilities and other characteristics, as well as behavior (team members who initiate ideas and support and praise others versus team members who interrupt and criticize others and propose misleading strategies). Pilot studies have found similar levels of motivation to accomplish the task, time on task, and problem-solving success among students interacting with a computer agent and students working in the same online environment but with a human partner (Rosen & Tager, 2013). Computer agents may also play the role of a nonpeer. For example, in Alelo Inc.'s program to teach foreign languages and assess students developing proficiency in a new language, students interact with a native-language-speaking avatar to carry out tasks such as negotiating between conflicting parties to resolve an argument (Soland, Hamilton, & Stecher, 2013).

Whether computer agents can be designed that will reliably mimic realistic conversational partners is not known. For example, research on AutoTutor, a computer program that converses with students using natural language, shows that the computer does not reliably detect and classify emotions (e.g., frustration, confusion, surprise), makes errors in interpreting the content of students' utterances (especially students' questions), and sometimes responds inappropriately due to misclassifying students' speech acts (Graesser, Rus, D'Mello, & Jackson, 2008). Limitations of a computer agent's communication facility may lead human participants to respond in unnatural ways, thus calling into question the validity of scores derived from human-computer interaction.

[24] See http://www.actfl.org/professional-development/certified-proficiency-testing-program/testing-proficiency

In conclusion, a number of questions remain to be answered, including the extent to which interacting with computer partners generalizes to interacting with human partners, whether using computer agents as partners produces groupwork experiences that are comparable from test taker to test taker, how many standardized group compositions are needed for generalizable scores, and how to select group compositions to represent the target domain of group compositions.

Variation Due to Role Assignment in Groupwork

Instructional research shows that role specialization can influence groupwork. To raise the level of discussion in groups, students can be assigned various roles such as recaller or learning leader roles to summarize the material and listener to ask questions, detect errors, and identify omissions in learning leaders' summaries (O'Donnell, 1999). Assignment of roles may influence group process. Schellens, Van Keer, and Valcke (2005) also found that assigning students the roles moderator, theoretician, summarizer, and source searcher in asynchronous online discussion groups produced more high-level communication about the task (e.g., testing and revising new ideas constructed by the group) than did group discussion without role assignment. Even in the absence of explicit role assignment, group members may assume specific roles that influence group dynamics, such as students positioning themselves as experts and novices and exhibiting asymmetrical teacher-learner interaction (Esmonde, 2009).

Recognizing that role assignment may influence group collaboration, PISA plans to include tasks that differ according to role structure: Some tasks will have symmetrical roles (every group member has the same role) and others will have asymmetrical roles (different roles are assigned to different group members, such as scorekeeper versus machine controller, Organisation for Economic Co-operation and Development, 2013).

Variation Due to Type of Task

There is increasing recognition, especially in research on managerial assessment centers, that the type of task may influence group processes and outcomes of groupwork. For example, role-play exercises, simulated interviews, and leaderless group discussions are designed to call on different groupwork skills (Howard, 2008) and may activate expression of underlying traits to different degrees and in different ways (such as extraverted test takers exhibiting influence more during leaderless group discussions than when giving oral presentations, Lievens, Chasteen, Day, & Christiansen, 2006). Indeed, reviews and meta-analyses of assessment center research show that ratings across different types of exercises are weakly to moderately correlated, even for the same groupwork dimension (e.g., Arthur, Day, McNelly, & Edens, 2003; Bowler & Woehr, 2006; Lance, Dawson, Birkelbach, & Hoffman, 2010).

The type of groupwork task may also change the *distribution* of group members' contributions within the group. One dimension of task type is the degree of structure, such as tasks with well-defined procedures and answers that can be completed by one person (called disjunctive tasks, Steiner, 1972) versus tasks with ill-structured solutions that cannot be completed very well by a single individual due to complexity or because no individual is likely to have all of the necessary expertise (Cohen, 1994; Cohen & Cohen, 1991). Chizhik, Alexander, Chizhik, and Goodman (2003) found that ill-structured tasks promoted more equally distributed participation among group members than did well-structured tasks.

Acknowledging that different types of tasks may require different groupwork skills, PISA plans to include different types of collaborative problem solving tasks that elicit different types of groupwork interactions and problem-solving behaviors. A possible typology of

tasks includes "(a) group decision making tasks (requiring argumentation, debate, negotiation, or consensus to arrive at a decision), (b) group coordination tasks (including collaborative work or jigsaw hidden profile paradigms where unique information must be shared), and (c) group production tasks (where a product must be created by a team, including designs for new products or written reports)" (Organisation for Economic Co-operation and Development, 2013, p. 22). The variety of task types needed to represent the domain of task types well is not yet known.

Variation Due to Task

Consistent with task variability in individual performance assessment scores (see Chapter III), group process and group performance may be quite variable even across tasks designed to be similar and to require similar competencies. Evidence of task variability comes from a variety of groupwork settings, such as military team simulations, simulation of medical-patient interactions, and simulation of teams in business and management. For example, Brannick, Prince, Prince, and Salas (1995) designed two simulated military air crew missions to be comparable in terms of group processes (e.g., assertiveness, decision-making, communication, adaptability) and performance expectations (e.g., performance of the designated pilot). Despite the careful matching of tasks, correlations between scores on the two tasks were low on both ratings of group processes and group performance. Few studies have estimated the number of tasks needed for dependable measurement of groupwork skills and group performance, although the results available suggest that the number may be very large (e.g., measuring medical students' clinical skills such as establishing and maintaining good [e.g., simulated] patient-physician rapport, Richter Lagha et al., 2012).

Variation Due to Occasion

Consistent with occasion variability in individual performance assessment scores, group process and group performance may be quite variable across occasions. Some studies show improvement in groupwork scores over time (such as improvement in students' negotiation skills from one session to the next, O'Neil, Allred, & Dennis, 1992), although the improvement may taper off over time (such as teamwork skill scores increasing over the first four fighter aircraft simulation missions and then remaining level after that, Mathieu et al., 2000). Other studies show instability in the relative standing of contributions of individual group members across occasions (such as Kenderski's, 1983, finding that some students exhibited high levels of help-seeking behavior on one occasion, while other students did so on another occasion; see also Webb, 1984). On the other hand, some evidence indicates that groupwork behavior may be more stable across time intervals within the *same* occasion (such as Meier, Spada, & Rummel's [2007] strong correlations for information pooling between time blocks among students working in dyads to diagnose psychiatric cases).

Variation Due to Type of Rater

As is the case for individual assessments, assessments with groupwork can use expert observers (rating live or recorded groupwork), peers, or test takers themselves as raters.[25] In the groupwork context, in contrast to individual assessments, peers are typically the other members of groups performing groupwork activities (e.g., members of triads of test takers in managerial assessment centers evaluating their peers' level of activity, persuasiveness, and clarity of communication in decision-making tasks, Shore, Shore, &

[25] It should be noted that other self-rating methods for gauging teamwork skills include questionnaires asking respondents to rate their own skills (e.g., I am a good listener) and multiple-choice situational judgment tests asking test takers to pick the best option for resolving hypothetical groupwork scenarios or pick the option that best represents how they would react (National Research Council, 2011). Because these measures typically are not directly tied to actual groupwork activities, we do not consider them further here.

Thornton, 1992). Although peer and self-ratings are less resource-intensive than expert ratings based on observations of groupwork (Dyer, 2004; Salas et al., 2005), lack of convergence with expert ratings is one reason why peer and self-ratings are unlikely to be used in high-stakes or summative assessments. Findings reported include (a) low to moderate correlations for cooperation, giving suggestions, accepting suggestions between observer and peer ratings of dyads flying simulated aircraft missions (Brannick, Roach, & Salas, 1993); (b) low correlations between medical students' self-reports of their behavior when interacting with standardized patients in a clinical performance assessment; and (c) experts' ratings of videotapes of the same encounters (Richter Lagha, 2013), and self-ratings, peer-ratings, and observer-ratings giving rise to different pictures of communication networks, and the centrality of specific individuals within them, in social network analyses (Bernard, Killworth, & Sailer, 1980; Kilduff, Crossland, Tsai, & Krackhardt, 2008; Kumbasar, Kimball Romney, & Batchelder, 1994).

Another issue regarding peer and self-ratings is their lack of independence. When team members rate each other, themselves, or their team as a whole on, say, contributions to the team's work, interactions with teammates, possession of relevant knowledge, listening ability, appreciation of different points of view, and conflict-resolutions skills (e.g., Loughry, Ohland, & Moore, 2007; Ohland et al., 2012; Taggar & Brown, 2001), they are themselves participants in the groupwork experience they are being asked to rate.

Variation Due to Type of Rating or Rating Scale

When rating the group as a whole or the behaviors of individual group members, multiple types of ratings or rating scales are available. Raters may code the presence or absence of specific events that are expected to take place at a particular point in time (e.g., providing information as required or when asked, asking for clarification of communication, verbalizing plans for procedures/maneuvers; Fowlkes, Lane, Salas, Franz, & Oser, 1994), the frequency of group processes (e.g., making contributions to groupwork discussions that refer to other group members' ideas; Rimor, Rosen, & Naser, 2010; or offering justified claims during argumentation; Weinberger, Stegmann, & Fischer, 2010), or the quality of specific behaviors observed (e.g., the effectiveness of conflict resolution, Fuchs et al., 1998; the quality of mutual support team members give each other during a critical phase of an exercise, Macmillan, Entin, Morley, & Bennett, 2013). Or raters may score groups or group members on general process dimensions (e.g., the quality of information exchange, communication delivery, supporting behavior, initiative and leadership; Smith-Jentsch, Cannon-Bowers, Tannenbau, & Salas, 2008; dialogue management, information pooling, Meier et al., 2007; collaborative problem solving, and communication, Taggar & Brown, 2001).

The evidence about the convergence of scores from different types of ratings or rating scales is mixed and is too limited to draw general conclusions. On the one hand, Macmillan et al. (2013) found substantial agreement between scores from analytic scoring of observable behaviors (e.g., rating of the team providing mutual support) and judges' overall ratings of team functioning on a scale from 1 to 5); on the other hand, Ohland et al. (2012, p. 625) reported modest correlations between scores on Likert-type and behaviorally anchored rating scales among peers who rated their teammates' contributions to the group's work (e.g., degree of agreement with statements that an individual *did a fair share of the team's work* vs. selecting the rating that best described an individual's behavior on a broad category such as *Does more or higher-quality work than expected; makes important contributions that improve the team's work; helps to complete the work of teammates who are having difficulty*).

Agreement Among Raters

As is the case for individual assessments, studies of observations of groupwork often report moderate to high agreement among raters. Such agreement, for example, has been shown in (a) rating live group interaction (e.g., the number of times fourth grade students provide explanations to other students when solving mathematics problems, Roschelle et al., 2009), (b) quality of support and problem-solving suggestions offered to others in the medical operating theater (Mishra et al., 2009), (c) rating videotaped groupwork (e.g., the frequency of informing other group members of critical information when flying simulated aircraft missions, Brannick et al., 1995), (d) judging audiorecorded groupwork (e.g., information exchange, communication delivery, and initiative and leadership in Navy teams, Smith-Jentsch et al., 2008), (e) rating online groupwork (e.g., the quality of argumentation in groups tasked with using theories about motivation and learning to understand and explain a student's poor performance in a mathematics course, Stegmann, Wecker, Weinberger, & Fischer, 2012), and (f) rating test takers' oral language proficiency during in-person or phone interviews (Ferrara, 2008). Moreover, research shows that a feasible amount rater training (e.g., 18 hours of practice rating and debriefing) can markedly reduce the magnitude of discrepancies between novice and expert raters' judgments of teamwork behavior (e.g., when rating the quality of communication, coordination, and cooperation in surgical teams, Russ et al., 2012).

Raters do not always agree even moderately, however, for reasons that may be specific to the groupwork setting. For example, raters evaluating conversations among pairs of examinees in a second language speaking test showed low agreement about examinees' language fluency and effectiveness (May, 2009). Raters interpreted the same conversations very differently. For example in asymmetric interactions (one examinee more talkative than the other), one rater may have perceived the less dominant partner as loafing (and downgraded that examinee as a result) while another rater may have perceived the same examinee as being unfairly suppressed (and upgraded that examinee to compensate for the unfair pairing). Such results show that raters need training in how to take into account the possible influence of different patterns of interaction among examinees (e.g., asymmetric, parallel, and collaborative, Galaczi, 2008) on their ratings of individuals' competencies such as oral language skills.

Automatic Coding and Scoring of Group Processes

Because human coding and scoring of group processes are very time-consuming and expensive, researchers are exploring automatic coding and scoring, especially in online environments that capture interaction automatically in data log files (e.g., capturing information flow in online forums to investigate networks of student participation, Zhang, Scardamalia, Reeve, & Messina, 2009). Approaches for automatically scoring content of interaction include classifying transcripts of conversations (either between students, or between student and computer agent) according to their similarity to known text (Foltz & Martin, 2008), and applying automatic speech recognition to analyze spoken communication (Johnson, 2010). Researchers are investigating automatic scoring of groupwork interaction. Measures include (a) how frequently or well members support each other (providing backup, correcting errors, Dorsey et al., 2009) or how frequently students refer to each other's contributions, formulate counter-arguments, and collaboratively apply scientific concepts to solve problems (Rose et al., 2008); (b) identifying the roles that individuals play at any given point in time (e.g., directing groupwork, asking questions of others, encouraging participation, Goodman et al., 2005); (c) scoring speaking and understanding of a foreign language when interacting with an avatar (a realistic, computer-generated human being) in simulated face-to-face conversations (Johnson, 2010); and (d) scoring students' communication about scientific methods (e.g., identifying flaws with news articles

or blogs about science) when interacting with computer agents during a computer game (Forsyth et al., 2013).

Emerging evidence about the agreement between human judges and computer programs when coding the text of communications is mixed. Rose et al. (2008) reported fairly high agreement between human coders and text classification algorithms for some group processes such as connecting arguments to create a group argumentation pattern, although not for others (e.g., referring to the contributions of their group partners). The higher reliability indices compare favorably to agreement among human raters using similar coding schedules (e.g., Schoor & Bannert, 2011). A major challenge for automatic scoring is how to construct classification algorithms that will transfer well between group discussion data for different topics and contexts (Mu, Stegmann, Mayfield, Rose, & Fischer, 2012).

An approach that bypasses the need to code ongoing interaction constrains the communication among group members to predefined instances of specific categories. Students choose from a menu of messages (derived from previous instances of unconstrained communication) that experts have judged to represent dimensions such as adaptability, communication, coordination (Chung et al., 1999), or building a shared understanding, problem solving, and monitoring progress (Rosen & Tager, 2013). The number of times a student sends messages in a particular category, such as decision making, forms the basis for the student's decision-making score. Menu-based interfaces may apply to fine-grained skills such as communicating with team members about the actions being performed, monitoring and repairing shared understanding, and prompting other team members to perform their tasks. For example, in the planned PISA assessment on collaborative problem-solving, test takers will be awarded points for taking specific actions such as asking the computer agent for the agent's point of view before implementing a plan (corresponding to the skill of building a shared representation) or monitoring whether the computer agent follows the plan as discussed (corresponding to monitoring and repairing the shared understanding, Organisation of Economic Co-operation and Development , 2013). How well constrained communication using predefined options maps onto natural communication without constraints remains to be investigated.

Variation Due to Random Coding Errors and Data Entry Mistakes

As is the case for individual assessments, unsystematic variation due to random coding errors and data entry mistakes can arise in assessments with groupwork. Here, however, a single error or unsystematic event can influence the scores of multiple test takers simultaneously and in different ways. For example, if one student's microphone malfunctions halfway through groupwork, reducing that student's contributions that are available for coding, the student may be scored too low while another student (whose contributions make up a larger share of the total as a result) may be scored too high. As another example, a rater who does not realize that two ideas are voiced by different students may credit one student for both ideas, inflating the score for one student and depressing the score for another. How effective usual strategies for minimizing the effects of errors on individual assessments, such as using multiple raters, multiple tasks, additional rater training, making imputations or adjustments for missing data, will be for assessments with groupwork remains to be investigated.

Random errors may also affect group-level scores and their reliability. Recent research in social network analysis suggests that the effects of unsystematic coding errors and data entry mistakes on the reliability of groupwork measures (such as an individual's centrality in a network) may depend on the particular kind of error in combination with both the particular set of relationships among team members and the particular role of the individual within the group. Wang, Shi, McFarland, and Leskovec (2012) examined the effects of different types of measurement error in network data, such as the omission of a relationship

between two members of a network, misspelling of a group member's name leading to the same individual being counted twice, or two individuals having the same name leading to them being considered to be the same person. They reported that some kinds of errors (mistakenly omitting a link between individuals) pose a bigger problem than other kinds of errors (mistakenly omitting an individual from the network), and the network measures for some types of networks (e.g., few clusters or subgroups) are more resistant to these errors than are other types of networks (e.g., many clusters or subgroups). Guided by their results, Wang et al. recommended targeted strategies for gathering additional data or cleaning the data to improve reliability, such as gathering additional data (or cleaning the data) for highly active individuals rather than for all individuals. This intriguing notion that it may be productive to collect additional observations for some individuals but not others, depending on their role in the network, may apply to groupwork scores other than network measures.

Implications for Validity and Reliability

All of the facets described above give rise to variability in assessment scores. An important question follows, namely, how to estimate the number and variety of conditions needed for dependable measurement of the target skills. Generalizability theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991), an important tool for understanding and estimating reliability and validity, can serve us well here. Generalizability studies estimate the magnitude of sources of error, and decision studies use that information to design a time- and cost-efficient measurement procedure.

Using the language of generalizability theory, the facets correspond to sources of error variation. The facets help define the possible *universe* of scores of interest. Specifically, the universe is defined by all combinations of conditions of the facets. Ultimately, we would like to know the universe score for an individual (e.g., an individual's ability to collaborate with others) or for a group (e.g., the quality of a group's solution to a complex problem) where the universe score is defined as the average of the individual's (or the group's) scores in the universe. For example, we want to know a student's ability to collaborate with others across all possible group compositions in which a student might work, all types of tasks, all occasions of test administration, all methods of coding and scoring, and so on. The question becomes, then: how well can we generalize from the observed score for an individual or for a group based on the particular instances of groupwork on an assessment to the universe score?

In considering how to answer this question, it is helpful to differentiate facets related to validity and those related to reliability. Facets that influence the meaning of the construct we term validity facets. If there is variation due to conditions of a validity facet, and the assessment includes some conditions but not others, the observed score may not represent the construct of interest. For example, we might be interested in test takers' teamwork skills whether they work with others during face-to-face interaction or in online interaction. If the two modes of communication generate substantially different scores, but only online communication is included in an assessment, generalization from the observed score to the universe score of interest will be severely compromised. As another example, if different types of rating systems (e.g., event-based ratings vs. ratings of general dimensions) do not converge, and only one rating type is used to rate observations of groupwork in an assessment, then the assessment scores may not generalize to the universe score of interest.

Efforts to standardize validity facets by purposively choosing conditions to include on the assessment (e.g., online interaction only) would produce scores that may generalize to a much more restricted universe than intended. Efforts to standardize validity facets may also change the nature of the generalization entirely. For example, if interacting with other

people and interacting with a scripted computer agent produce different scores, but the universe of interest is how well a test taker can communicate with other persons, standardizing an assessment to include only interaction with a scripted computer agent may produce observed scores that do not generalize to the desired universe at all. In both of these cases, the observed score does not fully represent the construct of interest. Estimating the variability in scores due to validity facets, such as through one or more generalizability studies, is a necessary step for making decisions about which conditions of a validity facet need to be included in an assessment to produce scores that generalize to the universe score of interest.

Other facets, which we term reliability facets, influence generalizability in a different way. Variability due to reliability facets influences the dependability of scores without necessarily changing (or restricting) the meaning of the construct measured. For example, suppose that a test taker's communication changes from one occasion to another (even when the group, the task, etc. are the same), but interest lies in generalizing over a wide range of occasions. Observing his communication on some occasions, but not others, may lead to questionable inferences to the universe score but should not affect the meaning of the observed score. The solution is to include as many occasions as needed so that the average score across occasions generalizes to the universe score of interest. Including reliability facets in generalizability studies can show how many conditions of reliability facets are needed for dependable measurement.

Despite knowing the many sources of measurement error that may influence scores from assessments with groupwork, we do not know the magnitude of the error from potential sources. Consequently, we do not yet know, for example, how many standardized group compositions, or role structures, or task types, or occasions are needed for dependable measurement of groupwork skills and performance. Designing and carrying out generalizability studies will help inform these questions.

Additional Issues Introduced by the Use of Groupwork in Assessments

Relationships Between Process and Productivity/Performance

As described throughout this chapter, assessments with groupwork may produce scores for multiple constructs, some related to process, others related to productivity or performance. These scores may not be highly, or even positively, correlated. For example, while some studies find significant, and even high, correlations between group processes (e.g., providing mutual support, information exchange, communication, team initiative and leadership) and quality of the team's performance or accuracy of decisions (Macmillan, 2013; Smith-Jentsch, Johnston, & Payne, 1998; Taggar & Brown, 2001), others have reported weak or nonsignificant relationships between similar processes and outcomes (Meier et al., 2007; Chung et al., 1999). Still others produce conclusions in opposite directions, such as the density of adversarial relationships being positively or negatively related to team performance (Baldwin, Bedell, & Johnson, 1997; Sparrowe, Liden, Wayne, & Kraimer, 2001).

One implication is that process and product/performance constructs are distinct and that measures of one cannot serve as proxies for the other. Another implication is that psychometric properties may vary for measures of process and product, and so may need to be investigated separately. For example, particular sources of measurement error may figure more prominently for some measures than for others (e.g., larger rater variation for measures of teamwork skills than for measures of group productivity). The optimal design for dependable measurement may, then, differ depending on the target construct.

Multiple Levels of Objects of Measurement: Individual and Group

Assessments with groupwork may yield scores at the individual level (e.g., a test taker's teamwork skills, ability to communicate about the subject matter, performance during or after groupwork) and at the group level (e.g., the teamwork functioning of the group, the quality of the group's product or performance). The possibility of multiple levels of objects of measurement for assessments with groupwork introduces complexities not found in purely individual assessments. One is that reliability of scores may differ from one level to another. The relevant sources of measurement error, and hence the optimal design for reliable measurement, may not be the same or may not function in similar ways when, for example, measuring individuals' teamwork skills and when measuring the functioning of the group as a whole. For example, raters may find it more difficult to rate individual group members' contributions to resolving conflicts than to rate the group's success in conflict resolution, giving rise to lower rater agreement for individual scores than for group scores.

Conventional approaches for examining validity may also yield different results depending on whether the object of measurement is the individual or the group. That is, expert-novice comparisons (e.g., Brannick et al., 1995; Fowlkes et al., 1994; O'Neil et al., 1992; Smith-Jentsch et al., 1998), predictions of future performance (e.g., Arthur et al., 2003; Meriac, Hoffman, Woehr, & Fleisher, 2008; Speer, Christiansen, Goffin, & Goff, 2013), and examinations of the dimensionality of groupwork measures (for example, through exploratory or confirmatory factor analyses, O'Neil, Chuang, & Baker, 2010; Taggar & Brown, 2001; or multitrait-multimethod analyses of the divergence of dimensions compared to the convergence of methods for measuring them, Brannick et al., 1993) may produce different results for individual and group scores. For example, giving suggestions and accepting suggestions may reflect teamwork skill dimensions that are more distinct (or separable) at the individual level than at the group level. These possibilities show that psychometric analyses need to attend to the particular unit of interest: individual, group, or both.

Another complexity is the statistical modeling issue arising from the non-independence of individuals within the group, especially when interest lies in producing dependable scores for individual examinees (on, say, a student's ability to collaborate with others, or the ability to engage in scientific argumentation). In collaborative settings, individuals' contributions are linked to, and dependent on, the contributions of other group members. Hence, the assumption of statistical independence of individuals' scores in conventional psychometric methods may not hold. New methods being explored for reliably scoring individuals' contributions to dynamic interactions during collaboration include dynamic factor analysis, multilevel modeling, dynamic linear models, differential equation models, social network analysis, intra-variability models, hidden Markov models, Bayesian belief networks, Bayesian knowledge tracing, machine learning methods, latent class analysis, neural networks, and point processes (von Davier & Halpin, 2013; National Research Council, 2013).

Another statistical modeling issue related to non-independence concerns the lack of independence from one group to another. Consider the desire to estimate the variability in individuals' scores across multiple group compositions. One way to gauge this variability is to observe the same individual in multiple groups that vary in terms of group composition attributes (e.g., ability level). A complexity arises when these groups have members in common beyond the target individual. As a consequence of the shared membership (which may be termed a multiple membership structure), groups are not independent. Multiple membership models (Goldstein, 2003), which have been developed for similar situations such as estimating school effects when students attend more than one school (which might occur when a student changes schools mid-year), will be helpful here.

Yet another complexity introduced by the use of groupwork concerns task design. In an individual assessment, all work on a task comes from a single test taker. In assessments with groupwork, in contrast, one, some, or all members of a group may contribute to the task. If the intent is to measure collaborative skills (whether at the individual or group level), task designers must attend to features of the task that may inadvertently reduce opportunities for participation or communication with others. For example, easily divisible tasks or large and complex tasks may encourage groups to divide up the work and assign different group members different portion to complete, resulting in largely independent, rather than interactive, work. Similarly, tasks that can be completed by one person may also inhibit interaction among group members, albeit for different reasons. The desire to measure collaboration at the individual level poses an additional challenge: designing tasks that are likely to involve all, not just some, group members. In sum, then, task developers must be sensitive to possible consequences of task design for the nature and distribution of test takers' interaction on the task.

Conclusions

This chapter shows the immense complexity involved in arranging groupwork situations on assessments and the possible consequences for measuring groupwork processes, products and performance. The previous sections describe some strategies for investigating and addressing the complexity, such as conducting generalizability studies to estimate the magnitude of sources of variation and using that information to make decisions about the design of assessments (e.g., the number of tasks or task types to be included, the number of different groupings to use for each test taker), or using avatars to represent group member attributes and behavior in an attempt to standardize groupwork experiences from one test taker to another or from one occasion to another.

Given the many sources of variation that potentially influence the measurement of processes and outcomes of groupwork, the number of conditions needed for dependable measurement may simply be too large, especially when the object of measurement is the student. Two alternative strategies for dealing with this issue are as follows:

- Shift the focus from estimating the proficiency of the test taker (or small group) to estimating the performance of the classroom or school. For example, even if the number of observations (e.g., tasks, groupings) is too small for dependable measurement of individual students (or of small groups), in some circumstances aggregating individual-level scores to higher levels may produce dependable measures of skills at the classroom or school level. As described by Brennan (1995), we expect reliability of higher-level units, such as schools, to be greater than reliability of lower-level units, such as students, when the number of students sampled within schools is fairly large and variability of school means is large relative to the variability of students within schools.
- Apply matrix sampling such that different groups within a classroom or school are assigned different conditions (e.g., different collections of tasks, different sets of group compositions). Matrix sampling may be an effective way to reduce the number of conditions per group and still maintain reliability at the classroom or school level. Allowing different groups to be assigned to different conditions may also make it possible to handle a major challenge in assessments with groupwork–systematically manipulating group composition. Rather than making systematic assignments of particular group compositions to particular groups, which may not be feasible, an alternative is to form groups randomly. Doing so may help assure that a large number of group compositions will be represented across the classroom or school, and that effects associated with particular group compositions will cancel out in the aggregate.

References

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125–154.

Baldwin, T. T., Bedell, M. D., & Johnson, J. L. (1997). The social fabric of a team-based M.B.A. program: Network effects on student satisfaction and performance. *The Academy of Management Journal, 40,* 1369–1397.

Baron, J. B. (1992). SEA usage of alternative assessment: The Connecticut experience. In *Focus on evaluation and measurement. Proceedings of the National Research Symposium on Limited English Proficient Student Issues (Washington, DC, September 1991)* (vol. 1, pp. 187–233). Washington, DC: United States Department of Education, Office of Bilingual Education and Minority Languages Affairs.

Baron, J. B. (1994, April). *Using multi-dimensionality to capture versimilitude: Criterion-referenced performance-based assessments and the ooze factor*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Barron, B. (2000). Achieving coordination in collaborative problem-solving groups*. Journal of the Learning Sciences, 9,* 403–436.

Bernard, H. R., Killworth, P. D., & Sailer, L. (1980). Informant accuracy in social network data IV: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks, 2,* 191–218.

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology, 91,* 1114–1124.

Brannick, M. T., Prince, A., Prince, C., & Salas, E. (1995). The measurement of team process. *Human Factors, 37,* 641–651.

Brannick, M. T., Roach, R. M., & Salas, E. (1993). Understanding team performance: A multimethod study. *Human Performance, 6,* 287–308.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement*, *32,* 385–396.

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

Chiu, M. M., & Khoo, L. (2003). Rudeness and status effects during group problem solving: Do they bias evaluations and reduce the likelihood of correct solutions? *Journal of Educational Psychology, 95*, 506–523.

Chizhik, A. W., Alexander, M. G., Chizhik, E. W., & Goodman, J. A. (2003). The rise and fall of power and prestige order: Influence of task structure. *Social Psychology Quarterly, 66,* 303–317.

Chung, G. K. W. K., O'Neil, H. F., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior, 15,* 463–493.

Cohen, B. P., & Cohen, E. G. (1991). From groupwork among children to R&D teams: Interdependence, interaction, and productivity. *Advances in Group Processes, 8*, 205–225.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64,* 1–36.

Collins, L. G., & Hartog, S. B. (2011). Assessment centers: A blended adult development strategy. In M. London (Ed.), *The Oxford handbook of lifelong learning* (pp. 231–250). New York, NY: Oxford University Press.

Connecticut State Board of Education. (1996). *Connecticut Academic Performance Test (CAPT) interdisciplinary assessment.* Hartford, CT: Author.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley.

Dorsey, D., Russell, S., Keil, C., Campbell, G., Van Buskirk, W., & Schuck, P. (2009). Measuring teams in action: Automated performance measurement and feedback in simulation-based training. In E. Salas, G. F. Goodwin, & C. Shawn Burke (Eds.), *Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches* (pp. 351–381). New York, NY: Routledge.

Dyer, J. L. (2004). The measurement of individual and unit expertise. In J. W. Ness, V. Tepe, & D. R. Ritzer (Eds.), *Advances in human performance and cognitive engineering research: Vol. 5. The science and simulation of human performance* (pp. 11–124). Bingley, United Kingdom: Emerald Group Publishing Limited.

Esmonde, I. (2009). Mathematics learning in groups: Analyzing equity in two cooperative activity structures. *The Journal of the Learning Sciences, 18,* 247–284.

Fall, R., Webb, N. M., & Chudowsky, N. (2000). Group discussion and large-scale language arts assessment: Effects on students' comprehension. *American Educational Research Journal, 37,* 911–941.

Ferrara, S. (2008). Design and psychometric considerations for assessments of speaking proficiency: The English Language Development Assessment (ELDA) as illustration. *Educational Assessment, 13,* 132–169.

Foltz, P. W., & Martin, M. J. (2008). Automated communication analysis of teams. In E. Salas, G. F. Goodwin, & C. S. Burke (Eds.), *Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches* (pp. 411–431). New York, NY: Routledge.

Forsyth, C. M., Graesser, A. C., Pavlik, P., Cai, Z., Butler, H, Halpern, D. F., & Millis, K. (2013). Operation ARIES!: Methods, mystery, and mixed models: Discourse features predict affect in a serious game. *Journal of Educational Data Mining, 5,* 147–189.

Fowlkes, J. E., Lane, N. E., Salas, E., Franz, T., & Oser, R. (1994). Improving the measurement of team performance: The TARGETs methodology. *Military Psychology, 6,* 47–61.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Karns, K. (1998). High-achieving students' interactions and performance on complex mathematical tasks as a function of homogeneous and heterogeneous pairings. *American Educational Research Journal, 35,* 225–267.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the first certificate in English examination. *Language Assessment Quarterly, 5,* 89–119.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London, England: Arnold.

Goodman, B. A., Linton, F. N., Gaimari, R. D., Hitzeman, J. M., Ross, H. J., & Zarrella, G. (2005). Using dialogue features to predict trouble during collaborative learning. *User Modeling and User-Adapted Interaction, 15,* 85–134.

Graesser, A., Rus, V., D'Mello, S., & Jackson, G. T. (2008). AutoTutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner. In D. H. Robinson & G. J. Schraw (Eds.), *Recent innovations in educational technology that facilitate student learning* (pp. 95–125). Information Age Publishing

Hoffman, B. J., & Meade, A. (2012). Alternate approaches to understanding the psychometric properties of assessment centers: An analysis of the structure and equivalence of exercise ratings. *International Journal of Selection and Assessment, 20,* 82–97.

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology, 1,* 98–104.

Johnson, W. L. (2010). Serious use of a serious game for language learning. *International Journal of Artificial Intelligence in Education, 20,* 175–195.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65*, 681–706.

Kenderski, C. M. (1983). *Interaction processes and learning among third grade Black and Mexican-American students in cooperative small groups* (Unpublished doctoral dissertation). University of California, Los Angeles.

Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free rider effects. *Journal of Personality and Social Psychology, 44*, 78–94.

Kilduff, M., Crossland, C., Tsai, W., & Krackhardt, D. (2008). Organizational network perceptions versus reality: A small world after all? *Organizational Behavior and Human Decision Processes, 107,* 15–28.

Kumbasar, E., Kimball Romney, A., & Batchelder, W. H. (1994). Systematic biases in social perception. *American Journal of Sociology, 100,* 477–505.

Lance, C. E., Dawson, B., Birkelbach, D., & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods, 13,* 435–455.

Lazaraton, A., (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing, 13,* 151–172.

Levine, J. M., & Moreland, R. L. (1990). Progress in small group research. *Annual Review of Psychology, 41*, 585–634.

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91,* 247–258.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*, 1–16.

Loughry, M. L., Ohland, M. W., & Moore, D. D. (2007). Development of a theory-based assessment of team member effectiveness. *Educational and Psychological Measurement, 67,* 505–524.

MacMillan, J., Entin, E. B., Morley, R., & Bennett, W. (2013). Measuring team performance in complex and dynamic military environments: The SPOTLITE method. *Military Psychology, 25,* 266–279.

Malone, M. E. (2003). Research on the oral proficiency interview: Analysis, synthesis, and future directions. *Foreign Language Annals, 36,* 491–497.

Maryland State Department of Education. (1994). *Maryland School Performance Assessment Program: Public release tasks.* Baltimore, MD: Author.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology, 85,* 273–283.

Mathieu, J. E., Tannenbaum, S. I., Donsbach, J. S., & Alliger, G. M. (2014). A review and integration of team composition models: Moving toward a dynamic and temporal framework. *Journal of Management, 40,* 130–160.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26,* 397–421.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *Computer-Supported Collaborative Learning, 2,* 63–86.

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93,* 1042–1052.

Mishra, A., Catchpole, K., & McCulloch, P. (2009). The Oxford NOTECHS System: Reliability and validity of a tool for measuring teamwork behavior in the operating theatre. *Quality and Safety in Health Care, 18,* 104–108.

Mu, J., Stegmann, K., Mayfield, E., Rose, C., & Fischer, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *Computer-Supported Collaborative Learning, 7,* 285–305.

Mulryan, C. (1992). Student passivity during cooperative small groups in mathematics. *Journal of Educational Research, 85,* 261–273.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28,* 483–508.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010a). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects.* Washington, DC: Author.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010b). *Common Core State Standards for mathematics*. Washington, DC: Author.

National Research Council. (2011). *Assessing 21st century skills.* Washington, DC: National Academies Press.

National Research Council. (2013). *New directions in assessing performance of individuals and groups: Workshop summary.* Washington, DC: National Academies Press.

Neuberger, W. (1993, September). *Making group assessments fair measures of students' abilities*. Paper presented at the National Center for Research on Evaluation, Standards, and Student Testing's Conference on "Assessment Questions: Equity Answers", Los Angeles, CA.

Ockey, G. J., Koyama, D., & Setoguchi, E. (2013). Stakeholder input and test design: A case study on changing the interlocutor familiarity facet of the group oral discussion test. *Language Assessment Quarterly, 10,* 292–308.

O'Donnell, A. M. (1999). Structuring dyadic interaction through scripted cooperation. In A. M. O'Donnell & A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 179–196). Hillsdale, NJ: Erlbaum.

Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., . . . Schmucker, D. G. (2012). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self- and peer evaluation. *Academy of Management Learning and Education, 11,* 609–630.

O'Neil, H. F., Allred, K., & Dennis, R. (1992). *Simulation as a performance assessment technique for the interpersonal skill of negotiation* (Technical report). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

O'Neil, H. F., Chuang, S. S., & Baker, E. L. (2010). Computer-based feedback for computer-based collaborative problem solving. In D. Ifenthaler et al. (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 261–279). New York, NY: Springer Science+Business Media.

Organisation for Economic Co-operation and Development. (2013, March). *PISA 2015 draft collaborative problem solving framework.* Retrieved from http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Pomplun, M. (1996). Cooperative groups: Alternative assessment for students with disabilities? *The Journal of Special Education, 30*, 1–17.

Popham, W. J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, *89*(2), 146–155.

Richter Lagha, R. (2013). *Accuracy of professional self-reports* (Unpublished doctoral dissertation). University of California, Los Angeles.

Richter Lagha, R., Boscardin, C. K., May, W., & Fung, C. C. (2012). A comparison of two standard-setting approaches in high-stakes clinical performance assessment using generalizability theory. *Academic Medicine, 87,* 1077–1082.

Rimor, R., Rosen, Y., & Naser, K. (2010). Complexity of social interactions in collaborative learning: The case of online database environment. *Interdisciplinary Journal of E-Learning and Learning Objects, 6,* 355–365.

Roschelle, J, Rafanan, K., Bhanot, R., Estrella, G., Penuel, B., Nussbaum, M., & Claro, S. (2009). Scaffolding group explanation and feedback with handheld technology: Impact on students' mathematics learning. *Education Technology Research Development*, *58*, 399-419.

Rose, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *Computer-Supported Collaborative Learning, 3,* 237–271.

Rosen, Y., & Tager, M. (2013). *Computer-based assessment of collaborative problem-solving skills: Human-to-agent versus human-to-human approach* (Research report). New York, NY: Pearson.

Russ, S., Hull, L., Rout, S., Vincent, C., Darzi, A., & Sevdalis, N. (2012). Observational teamwork assessment for surgery: Feasibility of clinical and nonclinical assessor calibration with short-term training. *Annals of Surgery, 255,* 804–809.

Salas, E., Burke, C. S., & Cannon-Bowers, J. A. (2000). Teamwork: Emerging principles. *International Journal of Management Reviews, 2,* 339–356.

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a "Big Five" in teamwork? *Small Group Research, 36,* 555–599.

Salomon, G., & Globerson, T. (1989). When teams do not function the way they ought to. *International Journal of Educational Research, 13*, 89–99.

Sampson, V., & Clark, D. B. (2011). A comparison of the collaborative scientific argumentation practices of two high and two low performing groups. *Research in Science Education, 41,* 63–97.

Saner, H., McCaffrey, D., Stecher, B., Klein, S., & Bell, R. (1994). The effects of working in pairs in science performance assessments. *Educational Assessment, 2*, 325–338.

SCANS. (1999). *Skills and tasks for jobs: A SCANS report for 2000.* Washington, DC: Author.

Schellens, T., Van Keer, H., & Valcke, M. (2005). The impact of role assignment on knowledge construction in asynchronous discussion groups: A multilevel analysis. *Small Group Research, 36,* 704–745.

Schoor, C., & Bannert, M. (2011). Motivation in a computer-supported collaborative learning scenario and its impact on learning activities and knowledge acquisition. *Learning and Instruction, 21,* 560–573.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.

Shore, T. H., Shore, L. M., & Thornton, G. C. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology, 77,* 42–54.

Slavin, R. E. (1990). *Cooperative learning: Theory, research, and practice*. Englewood Cliffs, NJ: Prentice-Hall.

Smarter Balanced Assessment Consortium. (2012). *Thermometer crickets: Grade 11 performance task.* Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/performance-tasks/crickets.pdf

Smith-Jentsch, K. A., Cannon-Bowers, J. A., Tannenbau, S. I., & Salas, E. (2008). Guided team self-correction: Impacts on team mental models, processes, and effectiveness. *Small Group Research, 39,* 303–327.

Smith-Jentsch, K. A., Johnston, J. H., & Payne, S. C. (1998). Measuring team-related expertise in complex environments. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (vol. 1, pp. 61–87). Washington, DC: American Psychological Association.

Smith-Jentsch, K. A., Salas, E., & Baker, D. P. (1996). Training team performance-related assertiveness. *Personnel Psychology, 49,* 909–936.

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013, November). *Measuring 21st century competencies: Guidance for educators.* Santa Monica, CA: RAND Corporation.

Sparrowe, R. T., Liden, R. C., Wayne, S. J., & Kraimer, M. L. (2001). Social networks and the performance of individuals and groups. *The Academy of Management Journal, 44,* 316–325.

Speer, A. B., Christiansen, N. D., Goffin, R. D., & Goff, M. (2013). Situational bandwidth and the criterion-related validity of assessment center ratings: Is cross-exercise convergence always desirable? *Journal of Applied Psychology, 99*(2), 282–295. doi: 10.1037/a0035213

Stegmann, K., Wecker, C., Weinberger, A., & Fischer, F. (2012). Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science, 40,* 297–323.

Steiner, I. (1972). *Group process and productivity.* New York, NY: Academic Press.

Taggar, S., & Brown, T. C. (2001). Problem-solving team behaviors: Development and validation of BOS and a hierarchical factor structure. *Small Group Research, 32,* 698–726.

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations* (Research Report No. RR-13-41). Princeton, NJ: Educational Testing Service.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds. & Trans.). Cambridge, MA: Harvard University Press.

Wang, D. J., Shi, X., McFarland, D. A., & Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks, 34*(4), 396-409. doi: 10.1016/j.socnet.2012.01.003

Webb, N. M. (1984). Stability of small group interaction and achievement over time. *Journal of Educational Psychology, 76*, 211–224.

Webb, N. M. (1993). Collaborative group versus individual assessment in mathematics: Processes and outcomes. *Educational Assessment, 1*, 131–152.

Webb, N. M., & Mastergeorge, A. M. (2003). The development of students' learning in peer-directed small groups. *Cognition and Instruction, 21*, 361–428.

Webb, N. M., Nemer, K. M., & Zuniga, S. (2002). Short circuits or superconductors? Effects of group composition on high-achieving students' science performance. *American Educational Research Journal, 39*, 943–989.

Webb, N. M., & Palincsar, A. S. (1996). Group processes in the classroom. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 841–873). New York, NY: Macmillan.

Weinberger, A., Stegmann, K., & Fischer, F. (2010). Learning to argue online: Scripted groups surpass individuals (unscripted groups do not). *Computers in Human Behavior, 26,* 506–515.

Wildman, J. L., Thayer, A. L., Rosen, M. A., Salas, E., Mathieu, J. E., & Rayne, S. R. (2012). Task types and team-level attributes: Synthesis of team classification literature. *Human Resource Development Review, 11,* 97–129.

Wise, N., & Behuniak, P. (1993, April). *Collaboration in student assessment*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Zhang, J., Scardamalia, M., Reeve, R., & Messina, R. (2009). Designs for collective cognitive responsibility in knowledge-building communities. *The Journal of the Learning Sciences, 18,* 7–44.

V. Modeling, Dimensionality, and Weighting of Performance-Task Scores[26]

Introduction 83
Models for Relating Item Responses to Underlying Constructs 84
Matching Construct Models to Different Types of Item Response Data 84
Modeling Dimensionality 85
Expert Judgment in Developing Models of Dimensionality 86
Empirical Methods in Developing Models of Dimensionality 87
Accounting for Dimensionality 88
Special Issues With Modeling Performance Task Scores 88
Research or Diagnostic Use of Response Models 89
Dimensionality 89
Confirmatory Analyses 89
Exploratory Analyses 90
More Complex Factor Analysis Approaches 90
Issues and Measures of Model Fit 90
Operational Use of Response Models 91
Weighting 91
Case 1: Essential Unidimensionality 92
Case 2: Single-Construct Dimension With One or More Nuisance Dimensions 92
Case 3: Multiple-Construct Dimensions 93
Summary of the Weighting Section 94
Conclusions 94
References 95

[26] Lauress Wise was lead author for Chapter V.

Introduction

In the previous chapters, we defined performance assessment and indicated types of knowledge and skills to be measured; described and exemplified a variety of performance tasks, responses, and scales; and identified important sources of measurement error. This chapter focuses on how we use performance task scores, along with responses to other test items, to determine an individual's or group's standing on the construct or constructs we intend to measure.

More specifically, we identify a range of issues that must be considered in developing and testing models of task and item response data:

- *Type of response data*. Performance assessments may involve a wide variety of different score scales. For example, a test item might be scored as present or absent, correct or incorrect (etc.) on a 0,1 scale. Or it might be scored on an ordinal scale from poor (0) to very good (3). Or it might be scored as frequency (or percentage or proportions) of a certain action. Or the item response measure might be continuous such as response speed.
- *Score scales*. Standing on the underlying construct is typically modeled using a continuous scale that may be bounded (as in portion of a domain mastered) or may stretch to infinity in either or both directions as is the case with speed or rate models and most item response theory (IRT) models. However, reporting results in terms of sets of discrete categories, as in learning progression models or simply performance level, is a potential alternative, particularly for large-scale assessments that may support performance tasks.
- *Dimensionality*. Performance assessment involves complex tasks that are, as nearly as possible, simulations or representations of complex, diverse real-world (criterion) situations. Multiple facets of knowledge, skill and performance go into successful task completion giving rise to possible multidimensionality in scores. Multidimensional models may be needed to account for both intended and unintended dimensionality introduced by performance tasks.
- *Local independence*. Local independence requires that the partial correlations between items, controlling for standing on the construct measured, is zero. Local independence is a necessary consequence of strict unidimensionality (Lord, 1980, p. 19). However, it may be, particularly with performance assessment, that local dependence results from groups of test items that have been contextualized in a single complex problem (connectedness—see Chapter II). The individual items may thus have an internal dependence upon one another based on the test taker's understanding of this common problem, even though there is no evidence of clear and interpretable secondary dimensions. Ignoring local dependence can lead to bias in estimating reliability (Wainer & Thissen, 1996).

This chapter is divided into three parts. In the first part we discuss different models for relating task and item responses to an underlying construct or constructs. In the second part, we address the use of statistical models for two purposes: (a) to examine the appropriateness of hypothesized models for a specific assessment and provide evidence to support the validity of score interpretations derived from the assessment; and (b) for measurements for substantive research. And in the last part we discuss operational uses of statistical models including different approaches for creating an overall score, when either the data or the construct exhibits multidimensionality.

Models for Relating Item Responses to Underlying Constructs

Item response models link item response data to an underlying construct, indicating the likelihood of different response patterns as a function of standing on the construct. They inform how estimates of the construct being assessed should be constructed from the item response data. Options range from simply adding up the number of correct answers or raw score points for polytomously scored items to sophisticated maximum likelihood or Bayesian estimates from a specified probability model.

A second important use of item response models is that they provide a basis for estimating the magnitude of measurement error in the score estimates (see also Chapter III and IV). Measurement error may result from random variation in the particular items presented to a test taker, to luck in guessing answers when the test taker does not know the correct answer, and to scoring variation when constructed responses are scored by humans. Error estimates are important for form design to answer questions such as how many items are needed to achieve a desired level of score accuracy. Models are also useful in form construction to show how the distribution of item difficulties and discrimination levels relates to precision at different parts of the score scale.

Matching Construct Models to Different Types of Item Response Data

A variety of models is needed to cover different types of evidence required to support score interpretations. In addition, a variety of score scales may be used to describe standing on the underlying construct or constructs. Evidence from performance tasks may be dichotomously coded, correct/incorrect or present/absent, as with multiple-choice or many short-answer items. In other cases, performance task responses may be polytomously coded, as with scored responses from essays and other more extended response items, or they may be continuously measured, as when response speed is used to assess fluency or distinguish between expert and novice performance.

On the construct side, we distinguish between continuous and discrete models. Within continuous models, there are bounded models and infinite models. A common *bounded continuous model* results from attempting to assess the percentage of a targeted domain the test taker has mastered. It cannot get any worse than none or better than all. Percentage correct scores, with a number of assumptions about how well the items represent the targeted domain, are a type of bounded measure. Classical test theory (CTT) models focus on expected number or percentage correct scores as the scale of interest. By contrast, most IRT models conceive of a trait on which test takers are (more or less) normally distributed,[27] with the trait extending out infinitely in both directions. No matter how good (or bad) someone is, there could always be someone better (or worse), although estimates of trait standing are typically truncated so that there are minimum and maximum reported scores.

Table 5-1 lists possible combinations of response data types and construct model scale types and suggests classes of probability models associated with each combination. This paper is not intended as a general textbook on item response models; rather we focus on models for combining responses to performance tasks with responses to other types of items.

[27] IRT models, themselves, do not require any specific distribution of test taker ability, but many of the programs that estimate IRT parameters do use a marginal estimation approach that assumes normality.

Table 5-1. Probability Models for Different Types of Response Data and Construct Models

Type of response data	Type of construct model		
	Continuous—bounded	Continuous—infinite	Ordered discrete categories
Dichotomous	Classical test theory	Various IRT models	Latent class models
Polytomous	Simple score-point models	Partial credit; graded response	Loglinear models
Continuous	Inverse normal or logistic transformations	Regression models, generalizability theory models	Logistic or discriminant analysis models

Performance tasks often generate polytomous or continuous item scores. Where scores are in some sense comparable from one performance task to the next, generalizability theory (G-theory) models (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) provide a useful approach to sorting measurement error into components relating to test taker by task interactions and also scorer main and interaction effects. Results from G-theory analyses can be used to design assessments with a sufficient number of performance tasks to meet precision targets and also to monitor the effectiveness of procedures for training and monitoring scorers (see Chapters III and IV).

Modeling Dimensionality

The probability models described above most commonly deal with a single underlying construct dimension. Each can be extended to cover situations where multidimensionality is required to account for either multiple construct dimensions or for extraneous (nuisance) dimensions in the response data stemming from item format or task specific factors. Most commonly, the continuous construct models address dimensionality through simple structure, where different items measure different dimensions with little or no overlap. Of course, multidimensional item response theory (MIRT) models cover situations where skills on multiple dimensions may be required to answer individual items and multivariate regression or discrimination models may be used with more continuous item response data. Dimensionality in discrete construct models can be addressed through categorical models that are only partially ordered (Rutstein, 2012).

The potential for multidimensionality in response data is greatly increased when measuring growth over time. For example, in measuring gains in student achievement from one grade to the next, multidimensionality may be introduced by differences in the content standards for each grade level and also by differences in the effectiveness of curriculum and instruction during the intervening period.

The issue of dimensionality is particularly acute for performance tasks, which are inherently complex, typically demanding multiple abilities, skills, and actions to complete them. Performance tasks are usually expensive to develop, administer, and score in comparison to multiple-choice items. Some may argue performance assessments are needed because these tasks measure something different from what multiple-choice items measure, something that cannot be well measured with simpler item formats. This argument will only hold if we believe that the target construct is at least to some extent multidimensional.[28]

The question of how many dimensions it takes to explain correlations among item or task scores is both conceptual and empirical. However, the nature of the dimensions may vary in important ways. First, is the target construct conceptualized as a single thing or does it

[28] Of course, there may be unidimensional constructs that we believe cannot be meaningfully measured by selected response items at all.

have multiple aspects that may relate differentially to performance on different items? Second, do item or task characteristics, unrelated to what we are trying to measure, lead to increased correlations among items from the same task or decreased correlations between items from different tasks? For example, it may be difficult to know whether differences associated with different item formats are related to important construct differences or simply reflect nuisance factors. In some cases, nuisance factors may be associated with minor dimensions in the construct of interest that test developers may choose to ignore. These are questions that should be addressed through model-based research prior to operational use of the performance assessment as described in the section on research or diagnostic uses of statistical models below.

Figure 5-1 provides an example of a complex model of dimensionality for an assessment with both performance task and selected response data. The boxes on the left indicate the constructs that the assessment sets out to measure. These include both specific content standards (or claims) and an overall factor that accounts for correlations among scores for the specific standards. The boxes to the right indicate task specific factors that account for extraneous (unrelated to the construct being measured) correlations among scores from a single task and general methods factors that account for extraneous correlations among different tasks or items using the same measurement method.

The development of models that account for dimensionality in the response data may be based on a combination of expert judgment and empirical analyses. Empirical analyses, using a combination of confirmatory and exploratory analyses are described in the next section on research and diagnostic uses of response models. Different approaches to developing and refining dimensionality models, particularly as they pertain to performance tasks are described next.

Expert Judgment in Developing Models of Dimensionality

Test developers build process by content grids for test development from the definition of the target constructs.[29] Test items or tasks are designed to cover one or more of the cells of the grid. Increasingly, test developers begin with specification of specific claims about the content or processes that test items or tasks are designed to support (Mislevy & Haertel, 2006). Dimensionality may be introduced by design—by the different claims to be supported.

An important first question when performance tasks are involved is whether they are predominantly targeted to cover different content or cognitive categories in comparison to selected response or other simpler items. If so, it should be possible to define construct factors that reflect the separation of content covered by different item formats. If separate dimensions corresponding to the hypothesized factors are confirmed, a plausible case can be made that performance tasks cover a unique part of the content/cognitive domain.

One issue here is that it may be difficult to distinguish method factors from construct factors if different construct dimensions are measured by different methods. The traditional approach is to construct the test so that multiple types of measures are used in assessing each construct dimension (Campbell & Fiske, 1959). Given a limited number of performance tasks in most assessments, a completely balanced design may not be feasible. However if analyses are limited to two or three hypothesized construct dimensions, it may be possible to include sufficient overlap of methods for some of the constructs to allow some separation of method and trait effects.

[29] Additional complexity has been introduced by the *Next Generation Science Standards* (NGSS Lead States, 2013) with three targeted dimensions: practices, crosscutting concepts, and disciplinary core ideas. Performance tasks may be designed to measure varying combinations of these dimensions.

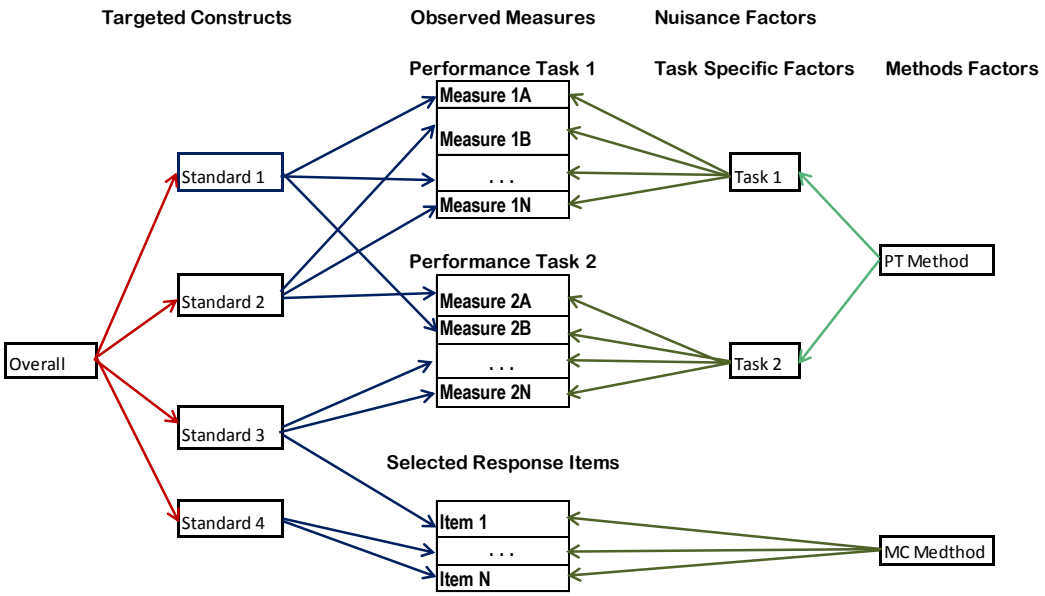

Figure 5-1. Example of a complex model of task and item data dimensionality. PT = performance task, MC = multiple choice, N = number of distinct items in a given performance task or the number of stand-alone multiple-choice items.

Empirical Methods in Developing Models of Dimensionality

Empirical methods are typically used in conjunction with expert judgment to examine dimensionality in scores. We distinguish and discuss two related approaches: inductive and deductive.

Inductive approaches for developing dimensionality models. When simple structure models (with each item related to a single dimension) are desired, it may be feasible to begin with a large number of hypothesized construct dimensions corresponding to detailed levels of the test blueprints and then use clustering methods (e.g., Tryon, 1939; Tryon & Baily, 1970) to combine dimensions with similar patterns of correlations into an increasingly smaller number.[30] At each stage, it is possible to test whether the reduction in dimensions results in significantly worse fit to the observed correlation matrix.

Of course, exploratory factor analyses are often used to develop dimensionality models. The number of underlying dimensions can be estimated empirically by various rules of thumb and the meaning of these dimensions can be inferred from the loadings of individual items on the individual factors after some appropriate rotation. (See Thompson, 2004.)

Deductive approaches for developing dimensionality models. The opposite approach is to begin with a single hypothesized dimension and then divide items into separate dimensions, either using analyses of residual item correlations or expert judgment. An alternative approach for separating distinct constructs would be to examine patterns of correlations with external variables. Suppose, for example, that a high school test is designed to measure readiness for college. It is possible that different items or item formats have

[30] Cluster analysis is similar to hierarchical factor analyses, albeit somewhat more primitive.

different patterns of correlation with different outcome measures, such as the need for remediation in a subject, freshman grades, and graduation rates. Different patterns of predictor weights for the different outcomes or even some form of canonical correlation analyses could be used to identify separate factors within the high school assessment.

Accounting for Dimensionality

As described above, multidimensionality that is not accounted for in the measurement model may lead to violation of the *local independence* assumption. Recall that local independence is the assumption that scores on different items or tasks are independent after controlling for person parameters. The presence of residual correlations among item scores means there is more work to be done to explain the factor or factors underlying test performance.

It is often the case that unanticipated nuisance dimensions are identified through empirical factor analyses as a result of such residual correlations. The residual correlations may be due to items from a common scenario, where understanding of the scenario may affect performance on all of the related items or possibly from familiarity and comfort with a particular item format in a way that is not related to the target construct.

A common approach for resolving local dependence issues relating to performance tasks is to combine items or scores associated with a common task into item bundles or testlets (Wainer, Bradlow, & Wang, 2007; Yen, 1993) and treat the resulting combination as a single, polytomously scored item. The main alternative to this approach is to add dimensions, such as task-specific factors, to the overall model to absorb the apparent dependencies. Another alternative is to simply ignore the violation of local dependence. Yen (1993) suggested that ignoring local independence violations may lead to underestimates of standard errors (the sum of correlated variables does not contain as much information as the sum of conditionally independent variables). Other research suggests that violation of local independence assumptions does not lead to significant bias in estimates of the underlying construct if there are not clearly distinct dimensions (Yen, 1984).

Special Issues With Modeling Performance Task Scores

Both classical test theory (CTT) and IRT models were honed on dichotomously scored items. Performance tasks, however, typically allow students to exhibit varying levels of proficiency on the construct being measured. generalized partial credit models (Muraki, 1997) or graded response models (Samejima, 1997) are typically used to relate standing on a continuous construct to the probability of achieving different score levels on polytomously scored items (see Osteni & Nering, 2006). Models specific to performance tasks that are continuously scored are less common, but a variety of common statistical models, such as regression and discriminant analysis models, can be applied to continuously scored performance tasks.

Connecting responses from performance tasks to the underlying construct involves some significant issues not associated with multiple-choice items. The first issue is that *human scoring* is often required to assign scores to student output from performance tasks. Item response models for performance tasks must take into account variation across human scorers and consistency in scoring over time. This is a situation where generalizability theory (GT) models are particularly useful in assessing sources of variation associated with test takers, items or tasks, and also elements of the scoring process.

A second issue is that, in some cases, performance task results may be measured on a continuous scale. One example is the time needed to respond to a set of mathematics fluency items. Time to a correct response is also frequently used as a measure to distinguish between novice and expert performance. Often a logarithmic transformation is

used to convert time measures from a zero to infinity scale to a scale that marches off to infinity in both directions and is thus more compatible with an unbounded score scale and a normal distribution of test taker scores. For more detail on combining response speed and accuracy information, see van der Linden (2007).

Another issue that arises with performance tasks is that there is often a desire to code the processes used to reach an answer in addition to scoring the answer itself. Examples include show-your-work tasks for mathematics problems and analyses of strategies for writing (e.g., outline, draft, and revision). This is another way in which performance tasks lead to a need for multidimensional models, since most people believe that process and outcome are at least somewhat different dimensions of performance. A related problem is that there is often not clear agreement that one process is universally better than others.

Research or Diagnostic Use of Response Models

Good testing practice requires that statistical models to be used for estimating construct scores and their precision be evaluated against empirical data prior to operational use. Of course, statistical models are also used in research based on operational testing that explores causal factors or consequences related to construct score estimates. The focus of this section, however, is on the former requirement for establishing evidence to support the hypothesized response models prior to operational use.

Two key issues must be addressed in evaluating alternative response models. The more specific issue is whether the model adequately accounts for the dimensionality of the response data. A more general issue is comparing the relative fit of alternative models to empirical data.

Dimensionality

In this section, we begin with a discussion of confirmatory analyses designed to model construct dimensionality. This is followed by a brief description of empirical analyses to test for the presence of nuisance factors, unrelated to the targeted construct. The final topic of this section concerns testing the overall fit of alternative models.

Confirmatory Analyses

Confirmatory analyses involve testing specific factor structure models against reasonable alternatives (see Harrington, 2009 for a fairly complete discussion; see also Cai, du Toit, & Thissen, 2011). Several different approaches to building and confirming or disconfirming construct models are open to those wishing to understand and document the importance of performance tasks. Models are typically built based on expert judgment, dividing the overall construct being measured into logical parts (e.g., different content standards for a given grade and subject). Multilevel models are often used to separate methods factors at one level from higher order content factors. Note that if models are developed empirically, as described above, an independent data set is needed for confirmatory analyses.

In most cases, researchers will analyze dimensionality based on data from a single point in time. For example, the Common Core State Standards assessment consortia currently only have data from extensive field testing to support dimensionality analyses. Ultimately, however, it will be useful to examine data from multiple points in time, either to confirm the stability of the hypothesized factor structure or to identify separate factors that may be highly correlated at one point in time but change differentially over time periods. For example, algebra and geometry skills may be highly correlated at any one point in time, but will change differentially if some students are exposed to more intensive or effective instruction in geometry while others receive more effective instruction in algebra.

Exploratory Analyses

Since unanticipated dimensions are not predictable, it is difficulty to build models of these dimensions in advance of data analyses. Exploratory factor analysis is a useful tool for uncovering unintended multidimensionality in assessments. A number of detailed considerations in exploratory factor analyses (what to use in the diagonals of correlation matrices, how to treat dichotomous variables, determining the number of factors, factor rotation, etc.) are well covered elsewhere (e.g., Harman, 1976 or Thompson, 2004). Exploratory factor analyses play an important role in identifying unanticipated dimensionality, but we have little to add to the existing literature on how best to conduct or interpret these analyses.

In the prior section, we described alternative models for relating performance task scores to an underlying construct. The input to exploratory analyses will include task scores constructed according to such models. A number of statistical tests can be used to determine the number of dimensions required to explain correlational patterns. If more than one factor is found, there are different approaches to rotating the factor structure so that the pattern of loadings of items on the rotated factors provide a basis for interpreting the meaning of the different factors.

More Complex Factor Analysis Approaches

A number of options might be considered in examining the dimensionality of performance-assessment scores. Of particular concern is that such assessments are likely to produce scores that are dichotomous, polytomous and possibly continuous. Options in factor-analytic approaches can be captured along two separate dimensions (e.g., Wiley, Shavelson & Kurpius, 2014; see Figure 5-2): (a) approach—exploratory or confirmatory—and (b) item-level model—linear statistical models or full information factor analytic models based on IRT.[31] Exploratory and confirmatory approaches differ as to whether a statistical model is specified prior to estimation, as is the case in a confirmatory approach (Kline, 2005). Full information (IRT) approaches are typically favored over linear methods in cases for which the item responses are not believed to be distributed linearly with each targeted factor (Bock, Gibbons, & Muraki, 1988), as might be expected from dichotomous and polytomous scores. In particular, full information approaches (Cai et al., 2011; Muthen & Muthen, 2010) can overcome some of the known problems with linear item-level factor analyses (e.g., sensitivity of correlations to similarities in distributions, such as distinct factors appearing for easy items and for hard items).

Issues and Measures of Model Fit

So, how do we evaluate the overall fit of a particular response model and how do we choose among alternative models? Usually we work to assess whether one model fits the data better than others. The pattern of responses across items and students is predicted by models with one or a few parameters per test taker and one to several parameters for each item or task. Fit statistics tell us whether the pattern of observed responses is likely or unlikely given optimal estimates of item and test taker parameters. When one model is a special case of another (for example, the one parameter logistic IRT model is a special case of the three parameter model), a likelihood ratio test may be used to test the extent to which additional model parameters significantly improve the fit of the models to the empirical data.

[31] It should be noted that factor analytic models could also invoke ordinal statistical models that use linear extraction methods on matrices of polychoric correlations. Likewise, nonlinear models could employ alternatives to the IRT methods we describe. We do not consider such models here.

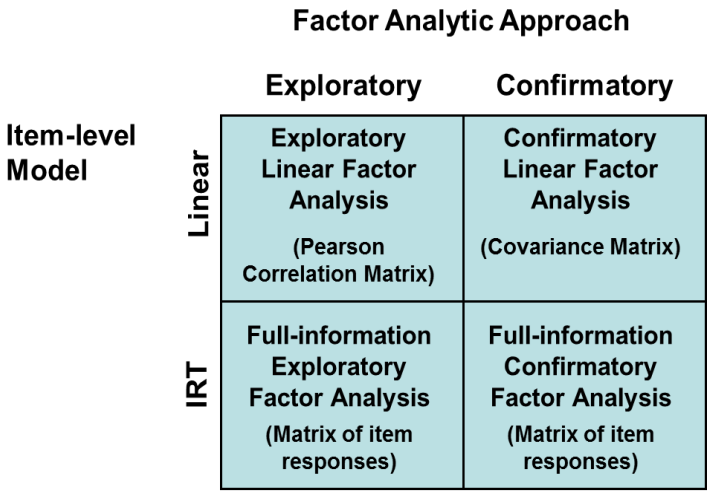

Figure 5-2. Alternative factor analysis approaches.

Note, however, that if we change the construct model to improve model fit, we run the risk of also changing the construct being measured, possibly in unintended ways. What is needed here is a combination of confirmatory factor analyses to model intended dimensions of the construct and exploratory factor analyses to discover any unintended dimension, or nuisance factors, that need also need to be accounted for. Recent advances in multidimensional item response theory (e.g., Haberman & Sinharay, 2010) may be useful in reporting results when multiple construct dimensions are confirmed.

Operational Use of Response Models

Response models have a variety of uses in supporting the operational estimation and reporting of assessment results. Uses such as calibration, scaling, and equating are discussed in detail elsewhere (see, for example, Lord, 1980) and not described further here. As noted above, the inclusion of performance tasks increases the likelihood that multidimensional models will be required. The key operational issue addressed in this section is the practical problem of how best to combine information from different item formats and construct dimensions into an overall estimate of the test taker's standing on the general construct of interest.

Weighting

In the first section of this paper, we discussed models for specifying response probabilities for performance tasks and other types of items in terms of standing on one or more underlying constructs. In the second section we described ways of collecting and analyzing evidence of the dimensionality of the underlying constructs. In this final section, we discuss approaches for combining performance task and other information into one or more indicators of the overall construct or constructs we seek to measure. We describe three general use cases based on the outcome of dimensionality analyses:

- Case 1: Essential unidimensionality
- Case 2: Single-construct dimension with one or more nuisance dimensions
- Case 3: Multiple-construct dimensions

In all three cases, it is important to be clear on the goal we are hoping to achieve in combining performance task and other information into some sort of weighted composite. For some purposes, in the third use case, it may be sufficient to report a profile of scores

across different underlying dimensions. However, an overall score is needed when important decisions are to be based on a characterization of the overall level of performance. We discuss three possible goals for developing composite scores when multiple dimensions are present:

1. Maximize the precision of the measure of the target construct, overall or at key decision points.

2. Maximize the prediction of one or more key criteria (e.g., indicators of college or career success).

3. Support one or more specific aspects of the theory of action for including performance tasks in the assessment in the first place (e.g., motivating changes in instruction to improve problem-solving skills; or engaging students more completely in the assessment tasks).

Case 1: Essential Unidimensionality

If there is only a single underlying dimension (or secondary dimensions are negligible), there might not be anything to combine. If our goal is to maximize the correlation with a criterion measure, we can accomplish that by maximizing the reliability of measurement of the underlying predictor construct. The true score correlation of a predictor with a criterion measure is attenuated by the (un)reliability of each of the measures. Thus the first two goals for combining performance task and other information, maximizing precision and maximizing predictive correlations, are the same if the predictor construct is unidimensional. Under IRT models, reliability may be maximized through maximum likelihood (or closely related Bayesian) estimates that fit responses to performance tasks and other items simultaneously. For classical test theory models, which are sometimes preferred for scoring transparency, Wang and Stanley (1970, p. 672) gave a general equation for the reliability of a composite based on the reliabilities and covariances of the component measures. Thus, for unidimensional data, reliability can be maximized using either IRT or classical test theory models, although the contribution of performance tasks to the overall score may not be sufficient to justify their time and expense under the reliability-maximization goal.

What if our goal is not just to maximize precision or prediction, but also to change instruction? It may be that the performance tasks and other items appear to measure the same thing at a particular point in time, but that changes in instruction could lead to differential improvement in one area relative to others over time. In this case, we may wish to establish explicit policy weights that communicate clearly the importance of the skills measured by the performance tasks. The key questions for those establishing policy weights is how much weight would the performance task scores have to be given for curriculum developers, teachers, and teacher trainers to take them seriously and how much precision might we willing to give up?

Case 2: Single-Construct Dimension With One or More Nuisance Dimensions

Nuisance dimensions for performance tasks may be associated with specific performance tasks, leading to significant test taker by task interactions, or may be a by-product of scoring,[32] leading to test taker by task interaction effects. Local item dependence may result from items associated with the same text or scenario being jointly dependent on a test taker's understanding of the context for a related set of items (Ferrara, Huynh, & Baghi, 1997; Ferrara, Huynh, & Michaels, 1999).

[32] Responses scored by the same scorer may be more highly correlated than responses scored by different scorers.

In factor analyses, unexpected correlations among items within a task may lead to spurious task factors not directly related to the overall construct being measured. Sometimes this problem is resolved by removing poorly performing or poorly fitting items or tasks. Often, a more sophisticated approach is needed.

The issue with nuisance dimensions is not so much how to weight them, as how to remove their effects as much as possible. One possible reason for such dimensions is that linear factor models are applied to response variables that are not linear, such as rate or speed measures (e.g., Wiley et al., 2014). By appropriately modeling responses such dimensions very well may drop out.

A second source of nuisance dimensions is item difficulty patterns. Adjustments to remove main effects of nuisance dimensions are common. For example, removing differences in overall task difficulties through equating is common. For scoring, scorer main effects (leniency) are typically removed through training and monitoring.

A third source of nuisance factors is due to interaction effects such as test taker by task. Such effects are not so easily removed. The usual approach is to average across a large enough number of tasks or scenarios to minimize the impact of task-specific factors on the overall measure. From a generalizability-theory perspective, the impact of test taker by task interactions decreases as the number of tasks in an overall measure is increased. The problem has been that it may be too costly or time-consuming to include very many performance tasks in a single assessment (see Chapter III).

Another approach to dealing with nuisance is to fit hierarchical factor models. Nuisance correlations may be explained by lower level factors, specific to individual tasks (or testlets). Reported scores would be based on higher-order factors that account for correlations across tasks.

Case 3: Multiple-Construct Dimensions

It may well be, as argued, that performance tasks are needed to assess constructs that are not well measured by multiple-choice or short selected-response test questions. If so, we will find that the performance space is at least somewhat multidimensional, leading to the issue of how to combine measures of individual dimensions into overall measures of achievement in a domain. It may also be that the performance tasks and/or selected response questions each cover multiple dimensions of achievement in a domain. In this case the question of weighting dimensions to arrive at a total score meeting the use criteria listed above must be addressed.

The first question, however, may well be whether to combine the measures at all. For example, the PARCC consortium is constructing separate end-of-course measures for Algebra I, Geometry, and Algebra II. Plans to combine these measures into an overall indicator of college and career readiness are still quite fluid. Some states might elect to use results from the Algebra II test alone as the indicator of college readiness in mathematics. Other states may choose to use the Algebra I test alone as a requirement for a high school diploma. Still other states may adopt a conjunctive model, requiring students to pass both the Algebra I and Geometry tests to receive a diploma.

A second example is PARCC's use of separate performance-based assessments (PBA) and end-of-year (EOY) assessments to create a summative measure of student learning. While models of the dimensionality of the two components of PARCC's summative assessment have not yet been tested, it is clear that an overall score will be required. Most likely, a compensatory model will be adopted, combining strengths and weaknesses across multiple dimensions into an overall measure. (See Ryan, 2002; Ryan & Hess, 1999; Winter, 2001;

Wise, 2010 for discussions of conjunctive versus compensatory rules for combining multiple measures and/or setting multiple performance standards.)

In most cases, however, an overall domain score is needed for school accountability, course evaluation, or reporting individual student achievement in a domain. Wainer and Thissen (1993) and Rudner (2005) describe empirical approaches to maximize score reliability (precision) or to maximize the correlation with a criterion measure. As pointed out by Wainer and Thissen (1993), these methods depend on the reliability of each of the measures being combined and that it may be possible to modify test lengths so that an equal weighting (or simple sum) of the separate components is optimal, either for reliability or for predictive validity.

A somewhat related goal might be to maximize judged alignment to a target domain. Often, key evidence for the validity of test score interpretations as mastery of a domain comes from alignment studies assessing the breadth and evenness with which the test items cover the domain. In such cases, component weights proportional to judgments of the relative importance of the parts of the overall domain covered by each component might be appropriate. More objective indicators, such as the amount of class or training time spent on particular topics, might also be used as a basis for defining component weights, but only if the skills measured by the individual components can be sufficiently tied to different parts of the curriculum.

A final approach to combining separate components is to consider consequences for instruction. Judgments, to be subsequently validated with follow-up studies, of the likelihood that specific weights will lead to positive changes in curriculum and instruction would form the basis for selecting component weights. Ultimately, however, we would want an independent measure of overall achievement to test or confirm the hypotheses that targeted changes to instruction did, in fact, lead to improved student achievement.

Summary of the Weighting Section

Multiple dimensions can be expected when complex performance is being assessed. Multiple dimensions associated with more specific areas of content within a domain are to be expected. Additional nuisance dimensions, unrelated to the target construct, must be eliminated or minimized to the extent feasible. Different objectives may be set for combining scores from construct-related dimensions into an overall measure. These include (a) maximizing score precision, (b) maximizing prediction of key criteria, (c) maximizing alignment to a target domain, and (d) ensuring desired consequent impacts on curriculum and instruction.

Conclusions

This chapter covered the (a) development models to relate performance task response data to the underlying construct or constructs we are trying to measure; (b) testing the assumptions of these models, particularly with respect to dimensionality; and (c) applications of the models to create an overall indicator for a particular use when multidimensionality is present. General conclusions/recommendations are as follows:

1. Models relating performance task scores to targeted constructs are useful in designing and constructing scores and in understanding the nature and extent of measurement error in estimating construct scores. Care is needed in selecting and testing such models.

2. Empirical verification of hypothesized construct dimensions is needed, particularly when performance tasks are combined with other item formats to cover complex target domains.

3. Exploratory analyses are needed to identify unintended or nuisance factors so that such factors can be accounted for appropriately in test design and score interpretation.

4. The purpose and intended uses of composite scores should be clearly specified in determining how best to combine scores from underlying dimensions or from individual tasks into an overall indicator.

References

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261–280.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modelin*g. Skokie, IL: Scientific Software International.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York, NY: Wiley.

Ferrara, S., Huynh, H., & Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large-scale performance assessment. *Applied Measurement in Education*, *10*, 123–144.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Measurement*, *36*, 119–140.

Haberman, S. J., & Sinharay, S. (2010). *How can multidimensional item response theory be used in reporting of subscores* (ETS Research Report No. RR-10-09). Princeton, NJ: Educational Testing Service.

Harrington, D. (2009). *Confirmatory factor analysis.* New York, NY: Oxford University Press.

Harman, H. (1976). *Modern factor analysis*. Chicago, IL: University of Chicago Press.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford.

Lord, F. M. (1980), *Applications of item response theory top practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hamilton (Eds.). *Handbook of modern item response theory* (pp. 153–164). New York, NY: Springer.

NGSS Lead States. (2013). *Next generation science standards: for states, by states*. Washington, DC: The National Academies Press.

Osteni, R., & Nering, M. K. (2006). *Polytomous item response theory models.* Thousand Oaks, CA: Sage Publications.

Rudner, L. M. (2005). Informed component weighting. *Educational Measurement: Issues and Practices, 20*(1), 16–19

Rutstein, D. W. (2012). *Measuring learning progressions using Bayesian modelling in complex assessments* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Ryan, J. M. (2002). Issues, strategies, and procedures for applying standards when multiple measures are employed. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 289–315*).* Mahwah NJ: Lawrence Erlbaum Associates.

Ryan, J. M., & Hess, R .K. (1999, April). *Issues, strategies, and procedures for combining data from multiple measures.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Samejima, F. (1997). Graded response models. In W. J. van der Linden & R. Hambleton (Eds.) *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer-Verlag.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: SAGE Publications.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington DC: American Psychological Association.

Tryon, R. C. (1939). *Cluster analysis: Correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*. New York, NY: Edwards Brothers.

Tryon, R. C., & Bailey, D. E. (1970). *Cluster analysis*. New York, NY: McGraw-Hill.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 73*, 287–308.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge, England: Cambridge University Press.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*, 103–118.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15,* 22–29.

Wiley, E. W., Shavelson, R. J., & Kurpius, A. A. (2014). On the factorial structure of the SAT and implications for next-generation college readiness assessments. *Educational and Psychological Measurement, 74*, 859–874.

Winter, P. C. (2001). *Combining information from combining information from multiple measures of student multiple measures of student achievement for school-level achievement for school-level decision-making: Decision-making. An overview of issues and approaches*. Washington, DC: CCSSO.

Wise, L. L. (2010, April). *Aggregating summative information from different sources*. Paper presented at the National Research Council workshop on best practices for state assessment systems, Washington, DC.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

VI. Challenges and Recommendations

Performance assessment is used to measure performance in education, work, and everyday life. Such assessment presents an activity or set of activities that requires test takers, individually or in groups, to generate products or performances in response to a complex task. These products or performances provide observable or inferable evidence of the test taker's knowledge, skills, abilities and higher order thinking skills in an academic content domain, in a professional discipline, or on the job. The psychometric challenges and recommendations/advice associated with performance assessment are summarized below.

Challenges

The psychometric challenges of performance assessment are often treated as the proverbial elephant in the room—the elephant is there, but no one wants to acknowledge or talk about it. The challenge, simply put, is that performance assessment involves complex, lifelike tasks and parallel real-life responses that can be intricate and lengthy and limited in number due to time and cost.

Standard psychometric models were developed for multiple-choice assessments, with many discrete test items that are scored dichotomously and that are organized into tests designed to measure one clearly defined construct. Performance assessment is complex and not so easily modeled, making it more difficult to confidently associate test-taker performance with a score or a performance category.

Some of the challenges associated with modeling performance assessment are

- Limited number of observations: Psychometric models work best when there are many of the same kind of observations of the same construct (e.g., many questions to assess reading comprehension). The time and cost associated with performance assessment puts a practical limitation on the number of observations that are possible.
- Complex and varied score scales: Performance assessments are not generally scored simply as either right or wrong. They might be scored using a rubric, or multiple rubrics, on scales that range from 0–3 or 1–6 or any other variant (percentages, error rate). They may also be scored using more unusual scales, such as the time required for a test taker to respond or some other process indicators. Further, the same performance assessment may result in multiple scores of different types.
- Human influence (raters): Performance assessments are often scored by human judgment. That is, raters are trained to read or observe student work and evaluate it based on the defined scoring criteria (e.g., rubrics). While a high-level of training and monitoring greatly helps to ensure rater accuracy, rater variation can introduce measurement error.
- Human influence (group members): Human influence can be particularly bothersome when the assessment is conducted in the context of groups. A student's performance on a groupwork skill (e.g., the ability to consider the ideas of others) is likely to be influenced by the behavior of the others in the group.
- *Connectedness*: The tools that psychometricians use to convert test taker performance to a score or category work best when various test questions/activities are unconnected (i.e., they satisfy the assumption of local independence). A performance assessment typically includes a set of activities, products, and item types that are designed to be connected. A complex performance assessment task,

for example, that requires a medical student to collect information and make a diagnosis may result in multiple scores based on many decisions or processes, but the scores would all be related to the same patient situation.

Dimensionality: Most psychometric models work best when an assessment measures one construct at a time (i.e., assumption of unidimensionality), so that the interpretation of the resulting score or performance category is clear. A performance assessment that requires a mathematics student to solve a complex multistep problem and then write about that process measures the student's ability to demonstrate multiple skills and therefore could confuse the interpretation of the results.

Recommendations

To address these challenges, at least partially, the following advice is given to those charged with making decisions about assessment programs:

1. Measure what matters. If the expectations for your students include the ability to apply knowledge to solve complex problems or deliver complex performances, work with your assessment advisors to explore the options for how to best measure those expectations. What gets assessed often commands greater attention.

2. If psychometricians or members of your technical advisory committee are concerned about the emphasis on performance assessment, listen to them. They are trying to advise you responsibly.

3. In cases where the construct of interest can be measured by multiple-choice assessment items, embrace that option. These items are both practical and reliable: (a) the format is familiar, allowing students to focus on the challenge of the content, (b) the items are administered in essentially the same way to each student on each occasion and produce comparable scores in either paper-and-pencil or computer-delivered mode, (c) the items are quick to administer allowing many observations in a given amount of time, and (d) the items can be scored accurately at very little cost in relatively little time.

4. Use complex performance assessment in cases where the construct of interest is important and cannot be assessed using multiple-choice items. Carefully define the construct(s) and select an assessment approach that requires students to demonstrate that construct in a way that is observable in their work. The necessary investment of dollars and time will need a clear justification.

5. Understand that measuring a complex construct accurately may require multiple observations, possibly under multiple conditions. Where consequences are attached to the assessment, it is critical that the inferences made are defensible. Given realistic time limitations, you may have to compromise on the length/complexity of the tasks in order to provide an adequate number of tasks under the required conditions.

6. If inferences are not needed at an individual student level, but rather at the level of the classroom or school, consider using a matrix sampling approach in which different students respond to different tasks. Increasing the number and variability of tasks may allow you to represent more thoroughly the construct of interest within reasonable bounds of time and expense.

7. Consider ways to enhance the performance assessment by combining it with other types of test items. This can be done with a mix of longer and shorter performance

tasks, as well as multiple-choice items, all assessing parts of the construct of interest in the situated context of the performance assessment. Using certain psychometric techniques (e.g., generalizability theory), it is possible to determine the number of tasks, multiple-choice questions, and raters needed to produce reliable scores.

8. Consider ways of making performance assessment part of the fabric of, for example, classroom instruction. If it is not taking time away from instruction, the investment of time for extensive or multiple performance tasks will not be as much of an issue. However, using the results in a summative fashion with stakes attached will require attention to standardization of conditions; steps need to be taken to ensure that the results represent the work of the test taker(s) being credited.

9. Consider various ways of combining results of performance assessment with multiple-choice assessment. Scaling a few (or a single) performance task(s) together with many multiple-choice items is likely to lead to a highly reliable overall score that is affected very little by the performance assessment. Treating the performance assessment(s) as separate from the multiple-choice assessment items and then combining them according to a policy-based weighting design will help make performance assessment scores count enough to warrant attention, but there may be a cost to the precision of the overall score.

10. Consider using performance assessment in the context of groups. Real-world performance often happens in the company of others, and the preparation of our students for that real-world challenge is getting increased attention. Keep in mind, however, that performance assessment in the context of groups is particularly complex, whether the work of the group as a whole is being assessed or the work of individuals within the group. Be careful about any inferences about individual students.

A Final Note

The psychometric group found that performance assessment poses interesting and important challenges that provide a rich agenda for research. In the foreseeable future, technological advancements may create new solutions as well as new challenges. Given new applications of performance assessment, including common assessments of K-12 students across multiple states, vast amount of response data will be created providing fertile ground for addressing that research agenda.

www.k12center.org

Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management at ETS has been given the directive to serve as a catalyst and resource for the improvement of measurement and data systems to enhance student achievement.