



TECHNOLOGY ENHANCED INNOVATIVE ASSESSMENT

Development, Modeling, and
Scoring From an
Interdisciplinary Perspective

Edited by Hong Jiao and Robert W. Lissitz

A VOLUME IN
THE MARCES BOOK SERIES



Technology Enhanced Innovative Assessment

**Development, Modeling,
and Scoring From an
Interdisciplinary Perspective**

A Volume in
The MARCES Book Series

Series Editors
Hong Jiao and Robert W. Lissitz
University of Maryland

The MARCES Book Series

Hong Jiao and Dr. Robert W. Lissitz , Editors

Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring From an Interdisciplinary Perspective (2018)

edited by Hong Jiao and Robert W. Lissitz

Test Fairness in the New Generation of Large-Scale Assessment (2017)

edited by Hong Jiao and Robert W. Lissitz

The Next Generation of Testing: Common Core Standards, Smarter-Balanced, PARCC, and the Nationwide Testing Movement (2015)

edited by Hong Jiao and Robert W. Lissitz

Value Added Modeling and Growth Modeling With Particular Application to Teacher and School Effectiveness (2015)

edited by Robert W. Lissitz and Hong Jiao

Informing the Practice of Teaching Using Formative and Interim Assessment: A Systems Approach (2013)

edited by Robert W. Lissitz

Computers and Their Impact on State Assessments: Recent History and Predictions for the Future (2012)

edited by Robert W. Lissitz and Hong Jiao

The Concept of Validity: Revisions, New Directions and Applications (2009)

edited by Robert W. Lissitz

Technology Enhanced Innovative Assessment

**Development, Modeling,
and Scoring From an
Interdisciplinary Perspective**

edited by

Hong Jiao

and

Robert W. Lissitz
University of Maryland



INFORMATION AGE PUBLISHING, INC.
Charlotte, NC • www.infoagepub.com

Library of Congress Cataloging-in-Publication Data

CIP record for this book is available from the Library of Congress
<http://www.loc.gov>

ISBNs: 978-1-68123-929-3 (Paperback)

978-1-68123-930-9 (Hardcover)

978-1-68123-931-6 (ebook)

Copyright © 2018 Information Age Publishing Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the publisher.

Printed in the United States of America

CONTENTS

1. The Virtual World and Reality of Testing: Building Virtual Assessments
Christopher Agard and Alina von Davier..... 1
2. Consistency and Validity in Game-Based Stealth Assessment
Valerie J. Shute and Gregory R. Moore31
3. Designing, Implementing, and Researching the Effects of Narrative-Based Assessment in Virtual Environments
Diane Jass Ketelhut and Brian C. Nelson 53
4. SimScientists: Affordances of Science Simulations for Formative and Summative Assessment
Edys S. Quellmalz and Matt D. Silbertglitt 71
5. Exploring the Psychometric Properties of Innovative Items in Computerized Adaptive Testing
Hong Qian, Ada Woo, and Doyoung Kim 95
6. A Noncompensatory Testlet Model for Calibrating Innovative Items Embedded in Multiple Contexts
Hong Jiao, Robert Lissitz, and Peida Zhan 117
7. Validity Issues for Technology-Enhanced Innovative Assessments
Brian Clauser, Melissa J. Margolis, and Jerome C. Clauser 139

vi CONTENTS

8. Principled Design and Development for Embedding Assessment for Learning in Games and Simulations <i>Kristen DiCerbo, Steve Ferrara, and Emily Lai</i>	163
9. Innovations in K–12 Assessment: A Review of CBAL Research <i>Randy Elliot Bennett, Rebecca Zwick, and Peter van Rijn</i>	197
10. ICT Literacy as a 21st Century Skill: Learning in Digital Networks With Agile Development Approaches <i>Mark Wilson, Kathleen Scalise, and Perman Gochyyev</i>	249
11. Visualizing the Process of Change in Learner Beliefs <i>Valerie J. Shute, Allen C. Jeong, and Diego Zapata-Rivera</i> ,.....	265
About the Authors.....	297

CHAPTER 1

THE VIRTUAL WORLD AND REALITY OF TESTING

Building Virtual Assessments

Christopher Agard and Alina A. von Davier

ABSTRACT

Traditional assessments are unable to address certain questions about why students perform as they do, and are unsuited to assessing important constructs like collaboration. The desire to create better assessments which can provide actionable evidence to improve students' skills and shape educational policies combined with recent advances in technology have led to the proliferation of virtual assessments. These assessments have the potential to provide the additional feedback and information needed, however, virtual assessments come with a number of psychometric and operational challenges that must be addressed if they are to realize their potential as assessment tools. In this chapter, we discuss the benefits of virtual assessments in relation to their traditional counterparts, psychometric and operational/logistic concerns regarding virtual assessments, and proposed solutions to mitigate or eliminate those concerns. We also provide brief descriptions of three recent virtual assessments—NAEP TEL's Wells Task, SimCity-Edu: Pollution Challenge! and the ETS Collaborative Science Assessment: the Tetralogue.

INTRODUCTION

The goal of this chapter is to discuss the benefits and challenges of virtual assessments as well as how to improve them. We use the term virtual assessment (VA) where other authors have used scenario-based and game-based assessment in order to include other types of assessment that can be conducted in a virtual environment and are designed to utilize features of that environment, such as computerized interactive assessments, virtual performance assessments, and computerized collaborative assessments.

The chapter commences by first offering a general definition of the terms “traditional assessment” and “VA.” Following that, we discuss the motivations for using VAs and follow this by defining two assessment components in detail: the assessment design and the assessment data. We describe the evidence centered design (ECD; Mislevy, 2011) concepts here, because we believe that it provides a useful conceptual framework in which to anchor a VA. Next, we discuss specific challenges to using VAs and potential solutions to those challenges. Finally, we consider some examples of VAs and we briefly describe how the design and data issues were addressed in these specific assessments.

We acknowledge that as a part of this discussion, serious consideration must be given to the operational and logistical concerns of assessments, if the assessments are to be implemented at any real scale and with any longevity. As such, we include, early in the section on challenges and potential solutions, a discussion of those concerns. We acknowledge that there are other challenges to building successful VAs, included issues of reliability, validity, fairness and test security. For this chapter we only address these issues as much as is necessary to discuss issues related to data (capture, management, analysis, and modeling), and operations and logistics. Those interested in issues of reliability, validity, fairness, and security are directed to the works of Mislevy et al. (2014).

Traditional assessments, in this chapter, refer to those assessments which do not use the virtual environment at all (“paper & pencil assessments”) as well as those which use the virtual environment to deliver items in a traditional format (multiple choice, MC, or free response, FR, item format) and produce similar types of data that paper and pencil assessments produce (computerized linear or even adaptive tests with traditional items). Computerized adaptive tests in use are tests that adapt the difficulty of the (MC or FR) items in real time to the estimated ability of the test taker and are examples of traditional assessments with features of the VAs. They are considered “traditional assessments” here because they often contain items that are either multiple-choice or constructed-responses and do not utilize data on interactions between the user and assessment beyond final responses. Examples of large-scale adaptive tests are TOEFL, GRE, and MCAT.

VAs include those assessments which capture additional information about user interactions with the assessment environment (process data), beyond final responses (outcome data). Many of these assessments may be scenario and game-based and are designed to make use of the virtual environment to deliver tasks that are closer to the real-life situations in which the constructs we wish to assess normally occur, but tasks which are not scenario or game-based could also be included here. We could think about collaborative problem-solving tasks or even imagine new versions of traditional assessments which would be designed to utilize process data. VAs need not be entirely composed of complex virtual tasks; they may still contain some traditional assessment items. Examples of VAs that include traditional item types exist in the NAEP TEL tasks (Keehner, Agard, Berger, Bertling, & Shu, 2014) which we discuss in the final section of the chapter.

Motivation

In recent years, technological advances have made it possible to use computers to capture rich data about the interactions of individuals with virtual environments and assessment developers have sought to leverage this capability to better understand the processes test takers employ to reach their final answers. This, at least on a large scale, had been impossible to do without the medium of the virtual environment. Now that the virtual medium has been made available for educational assessment, changes to how assessment is done are inevitable.

Traditional assessments have been criticized for their narrow focus with regard to the constructs that are being measured, but praised for their reliability and accuracy of their measures for constructs they have measured. This limitation was discussed in the MARCES presentation by A. von Davier (2014), but has been known for a long time (see, e.g., Katz, Martinez, Sheehan, & Tatsuoaka, 1998). A good assessment allows us to *know* whether or not a test taker has entered the correct final response. We infer, from the totality of a test taker's final responses and comparison to the correct response or to expert opinion (or comparative performance) in the area the assessment tests, how good an understanding a test taker has of the things the assessment covered. Traditional assessment performance, however, does not always match actual performance in the academic situations, and part of this dissimilarity may be linked to the dissimilarity in the context of the traditional assessment and the context in which knowledge is normally expected to be applied. By "normally", here we refer to nontesting situations.

In order to assess cognition from outcome data alone, we must assume that the final answer a test taker provides is in some way indicative of the underlying thought process that produced it. Even when responding to

an explicit question, those cognitive processes can vary from very careful reasoning based solely on the test taker's content knowledge to seemingly random responses (guesses) based upon unsystematic or arbitrary choices. Such variation in underlying cognitive processes is not necessarily reflected in the correctness of a test taker's responses, many of which might be dichotomous. Some attempts have been made in the educational measurement area to identify relevant factors that contribute to a correct or incorrect response and to identify students' misconceptions using traditional assessments. These studies employ cognitive diagnostic models and rely on experts' input about the cognitive attributes assumed to be needed for answering a specific question correctly (see Katz et al, 1998; Rupp, Templin, & Henson, 2010; von Davier, 2005). Often these attributes are highly correlated, and thus difficult to accurately estimate from the data. Hence, despite these efforts it is still challenging to provide actionable feedback to students based on traditional items.

VAs have several attractive qualities which may be used to provide adequate feedback, and enhance learning. They may create an organic link between teaching, learning, and assessment and can be embedded or linked into a curriculum and learning progression; include several teaching and testing strategies, such as direct or indirect knowledge sharing, through the use of included problems and access to resources; and also provide a natural environment for collaboration among test takers, either working in person or in remote teams (von Davier, 2014).

Perhaps most important to the intersections of assessment, cognition, and learning, VAs allow us to identify the strategies test takers employ and examine their problem-solving processes. Process data can be recorded and can be used to effectively reconstruct a test taker's actions during the assessment, allowing one to make inferences about aspects of that test taker's cognition based on those actions taken during the assessment. Those process data can be used to analyze the behaviors associated with final responses, in turn allowing us to form actionable hypotheses about how and why test takers provided the responses we see. In addition to this, process data can be used to identify concerns with other features of the assessment which may inadvertently impact outcomes for different types of test takers, for example, assessment design issues.

VAs can also allow the user to behave and interact more freely with the virtual environment, often involving the user in more or less realistic scenarios, and as a result, VAs are praised for being more engaging than traditional assessments (Mislevy et al., 2014). In light of their various benefits, VAs can be both a new medium for old assessments and a novel means of assessing otherwise difficult to assess constructs.

We have become very good at understanding and using traditional assessments. Advances in statistical modeling combined with compre-

hensive data collection and design have allowed us to better scrutinize performance of individuals and groups in these assessments. There is still much room and need for improvement, however. Traditional assessments give us very little information that we can use to determine why individuals do poorly, leaving us to assume that poor performance means poor understanding of the underlying construct. This can be particularly problematic for assessments with formative or policy implications. For instance, a state proficiency exam, which can only tell you that your students are performing poorly on the constructs that exam purports to test, does not give administrators and policymakers sufficient information from which they can make useful decisions on how to remedy the situation.

Environmental Benefits

Here we include a brief contemplation of an aspect of virtual assessment (VA)s that is tempting to consider when discussing the relative benefits of these tests over traditional one: the impact on the environment. In addition to the benefits discussed above, VAs also bring the possibility of environmental benefits by reducing costs associated with the development and administration process. Printed assessments include many environmental costs. While some costs associated with printed assessments may be reduced with reissue, many cannot, including for instance, environmental and financial costs associated with paper, ink, packaging, waste heat and water from servers, e-waste from new computers and tablets, plastic, cardboard, and fuel for delivery. Each accrues costs with each reissue. On the other hand, VAs have the benefit of mitigating some of these associated costs. The use of cloud services, energy efficient servers, and online delivery of content to existing infrastructure all offer opportunities for financial and environmental savings. It seems logical that the relative savings (environmental and financial) of VAs over traditional assessments would also increase with some direct proportionality to the number of additional administrations conducted. This logic holds if the delineation between VAs and traditional assessments one uses is the delivery medium of the assessments: a printed assessment has high environmental costs which accrue with each administration and virtually delivered assessment can mitigate their costs, as we have briefly argued. However, as we have defined VAs and traditional assessments in this chapter both can use a virtual medium of delivery and thus these potential benefits are not exclusive to VAs. Traditional assessments which use virtual delivery would therefore also benefit from a better entrepreneurial and environmental profile than their printed counterparts. Environmental benefits of virtual delivery should only be considered an advantage of VAs over printed traditional assessments.

EVIDENCE CENTERED DESIGN

Here we pause to describe evidence centered design (ECD) since the principles of this model for assessment design are relevant for the later discussion on addressing some of the challenges to designing virtual assessments. While the design for all assessments is the cornerstone of the work of content developers, cognitive scientists, and psychometricians, that design requires careful consideration of multiple elements and has to be rooted in the theory of the domain to be assessed. Furthermore, it must be closely tied to the test use. For VAs, the design is essential because the complex situational tasks in certain VAs increase the variance of the measurement of the construct of interest as well as construct-irrelevant variance, and the steps in the tasks have to be designed to elicit sufficient and meaningful information about the construct of interest. The evidence centered design framework (Almond, Steinberg, & Mislevy, 2002; Mislevy, 2011; Mislevy & Riconscente, 2006; Mislevy, Steinberg, & Almond, 2003) proposes a schema for the operationalized development of assessment(s) which works to keep the assessments grounded in a rationale appropriate to the domain(s) being assessed. ECD assigns assessment development activities to one of five distinct layers of the framework (Figure 1.1. Diagrammatic representation of the evidence centered design (ECD) framework for building assessments. Earlier assessment development actions occur in those layers closest to the left of the figure.). The foundational layer, *domain analysis*, refers to the gathering of relevant information about the domain being assessed and helps to identify appropriate kinds of problems and situations, necessary knowledge and skills, the ways in which information is typically represented, and how good work is characterized. The second layer, *domain modeling*, organizes the information gathered in the domain analysis to describe the relationships between the knowledge we wish to assess and the skills which demonstrate that knowledge. The third layer, *conceptual assessment framework* (CAF), is more operationally oriented and is where technical specifications of operational elements (e.g. measurement models, scoring methods and delivery requirements) are developed. The fourth layer, *assessment implementation*, deals with preparation of the operational administration and is where activities like authoring tasks, calibrating psychometric models, piloting and finalizing evaluation procedures, and producing assessment materials and presentation environments takes place, though some of these may already have begun prior to this phase. *Assessment delivery* is the final layer and is where developers deal with concerns regarding presentation of the actual assessment to intended test takers, relating scores to performances, and reporting the results. The ECD framework is the project plan for the building of assessments which result in more reusable assessment structures than would result from using

traditional assessment design frameworks—that reusability reduces the cost of developing these assessments.

In the initial design of traditional tasks, assessment content developers (content teams) and psychometricians need to ensure that the item works well, is fair, and measures the appropriate construct(s). In VAs, the same concerns remain, but are joined by concerns related to the new types of data (process data) we are capturing: “What are the relevant data that should be captured to support the claims we need to make about the test taker’s proficiency?”, “How do we capture them?”, and “How do we interpret them?” Addressing these concerns requires the skills of cognitive and learning scientists, data scientists (Press, 2013) together with the psychometricians. These conversations need to begin taking place as early as possible in the initial design of the task in order to better shape the assessment from its infancy. A better starting product, resulting from such conversations, coupled with practiced, responsible decision-making about changes, can reduce the overall number of changes made to a VA, thereby reducing the number of development iterations an assessment goes through, and its total cost.

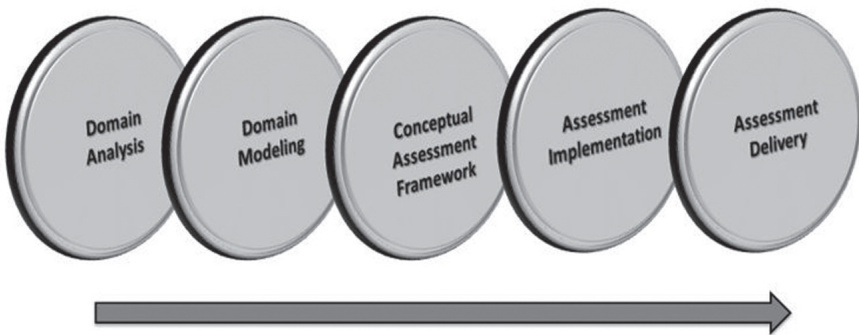


Figure 1.1. Diagrammatic representation of the evidence centered design (ECD) framework for building assessments. Earlier assessment development actions occur in those layers closest to the left of the figure.

TYPES OF DATA COLLECTED IN VAs

As indicated in the ECD framework, in order to be useful, the assessment data that are captured must be both *meaningful* and *sufficient*, and the assessment tasks must contain *relevant factors*. *Meaningful data* here are defined as process data and outcome data, but in some instances they may

include additional types of information, such as that from background questionnaires, pre- and posttest data, and so on. In order for the data to be *sufficient*, the series of actions/discourse for process data must be long enough to allow for appropriate analyses, and the number of items and participants for all types of data collected in the assessment must be statistically adequate. *Relevant factors* are those assessment or task features which elicit information about the test taker's behavior to support claims about the test taker's ability; these factors may impact the test results—for example, length or difficulty of the task, the type of task (e.g., collaborative or not). The relevant factors and the features that make the data meaningful are identified by the design team in the first stages of the ECD. There are a number of types of data produced by VAs. We discuss some of these data types below.

Outcome Data

Outcome data are the final responses to individual items or steps in a task that a test taker produces. Sometimes the outcome data are also called product data or responses. The outcome data are collected through the evaluative scoring throughout the assessment. For example, an individual's final responses during the assessment can be scored as *correct* or *incorrect* by a human rater or an automatic scoring engine. Pretests or posttests, if available, also result in individual outcome data. If either of these tests is available, then the test scores that contain information about the test-taker ability can be corroborated with the information contained in the actions scored throughout the virtual task.

Process Data

Process data refer to the additional information collected about user's interactions with the assessment environment. They can be used to reconstruct the specific behaviors a user takes within an assessment. They are also not unique to VA, although the medium of VA allows for more precise and rapid collection of larger amounts of process data than is feasible through manual collection. Timing information is also a type of process data that is readily available in most computerized assessments. Knowing how much time individuals spent on various items or parts of items can be useful in analyzing the motivation of the test taker, the speededness of the test, fatigue, and so on.

Background Data

Background data are collected via a background questionnaire (BQ) and provide demographic information about a given test taker. These data can be used together with the process and outcome data to support the claims of measurement invariance of the test results or to identify group strategies, for example.

Multimodal Data

Multimodal data (MMD) are multivariate process data that are useful for measurement and validity and can be used to augment the assessment data. MMD can include face-tracking, visual, audio, magnetic resonance imaging (MRI), and computed tomography (CT) data among others. The most recent technological advances make possible the collection of various biometric data via sensitive and accurate sensors like cameras, Kinect (www.xbox.com/en-US/kinect), and so on. These data can be used to provide feedback on how the content of the test is used, how the test takers interact with the virtual environment, whether they are motivated, engaged, and whether the cognitive load is in sync with their behavior.

Simulated Data

Simulated data are data created using models that are intended to mimic a given complex system with reasonable precision. According to Sokolowski and Banks (2009), the techniques of modeling and simulation associated with certain types of VAs have the potential to facilitate analyses of complex systems by providing relatively inexpensive data which can help us to understand interactions among those systems. Simulated data sets may be constructed to create a true state of the system that can be used to evaluate different statistical models. Examples of such systems are agent-based systems, computational models, and so on. Modeling and simulation capture the properties of a system in a computer and ask questions about that system. Simulations help develop the ability to (a) meaningfully simplify a complex problem, (b) capture the problem in a model, (c) describe the model in a computer language, (d) collect meaningful input data, (e) execute the model over time, (f) obtain and analyze results and test different statistical models, and (g) make inferences about a potential solution to the problem.

Collaborative (Problem-Solving) Data

Collaborative problem-solving (CPS) data and data from interactions of test takers result from tasks in which two or more individuals use individual and shared knowledge, skills and effort to create a solution to a given problem. CPS data are not unique to VAs, but can be collected as part of a VA, potentially enriching the understanding of interactions in a collaborative context. They are distinct from other process data in that they provide information about the interactions of multiple test takers relative to one another and the assessment environment during the collaboration.

CHALLENGES AND POTENTIAL SOLUTIONS TO BUILDING SUCCESSFUL VAs

Despite their engagement, motivation and potentially adaptive qualities, many VAs in education have been criticized for their error in measuring targeted constructs. They sometimes rely on poor or unidentified statistical models resulting in poor results, or are subject to too much construct-irrelevant variance. They often lack accuracy and proof of replicability and there is a paucity of work demonstrating good transfer of knowledge and other measures of validity. See the meta-analysis, “Our Princess is in Another Castle” (Young et al., 2012). There is, however, some evidence from the field of medicine that demonstrates the validity of various forms of simulation based assessments (McGaghie, Issenberg, Petrusa, & Scalese, 2010).

In order to overcome the major shortfall of VAs—error in measuring intended constructs—, it is important both to identify appropriate constructs and to develop quality measures of those constructs. To do this, interdisciplinary teams will need to make use of ECD principles to develop carefully designed studies and assessments. Those studies and assessments will need large representative samples which can be analyzed using appropriate and robust models, and will need to include (predictive) validity and fairness studies.

One problem, however, is that it is expensive, difficult, and time-consuming to collect and code data. This could be mitigated by leveraging tools such as crowdsourcing via Amazon Mechanical Turk (www.mturk.com) and simulation and modeling, thus reducing costs in the domain modeling stage of the ECD framework. Crowdsourcing web-services are useful for collecting large representative samples quickly and at potentially low costs. Modeling and simulation could also be employed to generate data rapidly, thereby reducing the need to find participants for certain studies (Bergner, Andrews, Zhu, & Kitchen, 2015; Sokolowski & Banks, 2009).

In the rest of this section we discuss the considerations needed around the data management and psychometrics, and the operational logistics that tie it all back to ECD and the data types described earlier.

Data Capturing and Data Management

One of the advantages of the VAs is that all of the data from the interactions between the test takers and the VA environment can be captured in log files. Log files are running time-stamped transcripts of user interactions with—and, in the case of virtual collaborative tasks, within—the assessment environment recorded in some structured file (e.g., xml or JSON).

Complex tasks in computerized educational environments provide additional measurement challenges, as relevant features of student performance must be extracted from the log files that are automatically generated as students work. Although process data from the log files are more comprehensive and detailed than most other forms of assessment data, the inclusion of such fine-grained detail presents a number of problems for analysis. Log files contain large quantities of information, typically consisting of thousands of pieces of information on each subject, with a single subject able to generate over 3,000 actions in just half an hour (Chung et al., 2010). The data are at such a small grain size (e.g., “selected Tool A,” “moved one space to the left”) that there is often no known theory to help identify precisely which actions are important to the construct being measured. There is often little overlap in the actions produced by one subject and the actions produced by a second subject making the log data large and sparse at the same time. The data are also noisy, in that irrelevant actions can vastly outnumber relevant ones and relevant actions are not usually identifiable a priori, and are so large that it is prohibitively costly to examine the data by hand. Pattern identification and matching methods from a data mining tool box can be used on the data from the log files to identify key performance features which correspond to strategies or classes of competencies.

In the example in Figure 1.2, one can begin to understand one of the challenges of raw log file data. They are not very easily interpreted by the human eye. The log file excerpt in Figure 1.2, reports the beginning of a task. Here we see that, in response to a user (or administrator) entering a task, the system immediately initializes an identified scenario and changes the slide the user sees. A second later, the system changes the screen again and not until 10 seconds later does the user first take action (changing the screen). Information is also provided about where in the task the action takes place, as well as that action’s beginning and ending.

```

<?xml version="1.0"?>
- <gamelog date="5/29/2014 4:59:31 PM">
  - <scenarios>
    <scenarioNo>-1</scenarioNo>
  </scenarios>
  - <scene>
    <sceneNo>-1</sceneNo>
    - <action>
      <actionName>Initialize</actionName>
      <actionTime>5/29/2014 4:59:31 PM</actionTime>
      <actionBy></actionBy>
      <actionTo>-1--1</actionTo>
      <actionResult>Started new session</actionResult>
      <slideName>Unknown</slideName>
    </action>
    - <action>
      <actionName>Changed slide</actionName>
      <actionTime>5/29/2014 4:59:31 PM</actionTime>
      <actionBy>System</actionBy>
      <actionTo>Players</actionTo>
      <actionResult>Active Slide: 000 - INIT EVERYTHING</actionResult>
      <slideName>Unknown</slideName>
    </action>
    - <action>
      <actionName>Changed slide</actionName>
      <actionTime>5/29/2014 4:59:32 PM</actionTime>
      <actionBy>System</actionBy>
      <actionTo>Players</actionTo>
      <actionResult>Active Slide: 001 - Start Slate</actionResult>
      <slideName>INIT EVERYTHING</slideName>
    </action>
    - <action>
      <actionName>Changed slide</actionName>
      <actionTime>5/29/2014 4:59:49 PM</actionTime>

```

Figure 1.2. Example of logfile data.

Data Dependencies and Psychometric Considerations

VAs that are based on complex scenarios and games or involve collaborative problem-solving present real challenges to traditional assessment techniques. The data these assessments generate tend to be interactive, interdependent, and dynamic in nature (Blech & Funke, 2005, 2010; Klieme, 2004). Their complexity often results in a wider range of suitable solutions, less control over how test takers reach those solutions, and more complexity in tracking progress toward those solutions. These very features create contexts for assessment which more closely mirror those in which the demonstration of the knowledge we wish to measure would naturally occur, but require different psychometric approaches than those in traditional assessment.

To discuss the psychometric considerations around the VAs we focus on virtual collaborative assessments because they seem to be the most complex in terms of data types and possible methodologies. The psychometrics for other VAs can be subsumed under the discussion here.

Analyzing data from a VA that includes collaborative problem-solving (CPS) tasks involves several modeling aspects that are not encountered in traditional assessments: (a) the data size and granularity (now we must also consider the hundreds or thousands of test-taker actions per assessment as opposed to only the final responses those actions produce), (b) the multidimensionality (both in terms of the individual's skills needed to solve a task and of the multiple time series of data, that is, the number of people in a team who may have different profiles of those skills), and (c) dependent data (this is problematic for traditional psychometric analyses which often are designed to work only with independent data).

Clearly, in assessments which involve a degree of collaboration, the local independence assumptions (LIA) needed in traditional psychometric analysis, such as in an item response theory (IRT) model, do not hold due to interaction among collaborators impacting behaviors. They also may not hold within tasks due to the dependence of the complex items within tasks, or within individuals due to effects of earlier behaviors on subsequent behaviors. This complex data dependency is illustrated in Figure 1.3. The data from a dyadic interaction consist of two observable time series of actions (X_1, \dots, X_n) and (Y_1, \dots, Y_n) from two test takers, AA and BB, with (potentially multidimensional) latent ability distributions A_1, \dots, A_n and B_1, \dots, B_n , respectively, that evolve over time as a function of the interactions between AA and BB and as a function of their states at time $t = 1, \dots, n$. It is obvious in this situation that the independence assumption neither holds over items (the responses X_{t+1} are dependent on A_{t+1} , which in turn depends on A_t and X_t and on Y_t and B_t) nor over people, since X_t and Y_t are not independent.

If the task is not collaborative, then the data dependencies may be showed by using a modified version of Figure 1.3. to represent a simpler VA without the second set of latent and observable variables. In that case, we would still need to deal with the dependencies among items due to the multidimensionality of the construct and due to the time dependence.

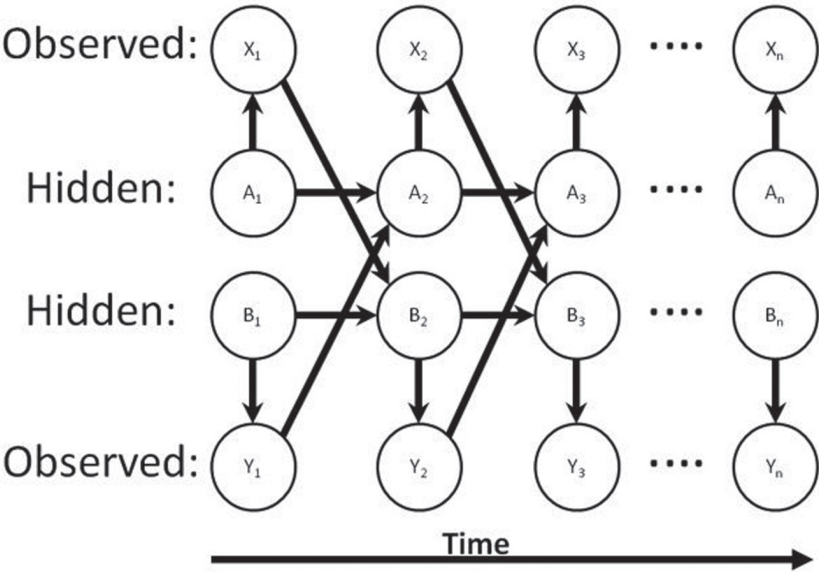


Figure 1.3. Diagrammatic illustration of data dependencies in dyadic collaborative problem-solving tasks.

Next we discuss static and dynamic features of data dependencies.

1. Static Features

Static features refer here to the data features at specific time points and are as follows:

- a. The item responses may be dependent on each other due to the multidimensionality of the construct.
- b. Item-within-task dependence may characterize tasks.
- c. Team dependence may exist: the individual's performance may depend on who is on the team, for example, some teams may have a better cohesion than others.

2. Dynamic Features

Dynamic features refer to the data features over the course of the VA and are as follows:

- a. The item responses from items within a complex task may depend on the responses to the previous items over a longer period of time.

- b. The item responses from collaborators may depend on the collaborators' previous actions.¹
- c. If there are numerous teams that change over time we may also conceivably have dynamic team dependence.

Next we will talk about static models versus dynamic models in order to summarize the existent statistical models available for addressing these types of dependencies mentioned above.

Static models that address the issues from 1.a and 1.b above, such as Bayesian network analyses (Bayes nets) account for item-within-task dependence. Bayes nets start by characterizing aspects of students' knowledge and skill in terms of a possibly vector-valued student model variable, A , as in Figure 1.3, and aspects of their behavior in terms of possibly vector-valued observable variables, X . Conditional probability distributions $P(X|A)$, obtained through theory, expert opinion, empirical data, or some combination of these, characterize how performance depends on knowledge and skill in task situations. Letting the prior probability distribution $P(A)$ denote the assessor's belief about a student's knowledge and skill at a given point in time, observing X leads to an updated posterior probability distribution $P(A|X)$ by Bayes' theorem.

Other models that can be used to account for these types of data dependencies are multidimensional IRT (MIRT) models (Yao, 2008), testlet models (Wainer, Bradlow, & Wang, 2007), and bi-factor models (Gibbons & Hedeker, 1992).

Static statistical models that may address the issues from above and account for test takers' dependence (team/group dependence) may include multilevel models, hierarchical segmentation, nonparametric exploratory models, latent class analysis (LCA)/neural network algorithms (Polvichai & Khosla, 2002). For example, latent class analyses (LCA) and neural networks algorithms model systems of variables that remain constant over time; multilevel models take into account clustering in the data such as team members with their teammates (or as repeated observations within individuals).

Dynamic models that address 2.a and 2.b above, and account for multidimensionality and item dependence over time, often assume a continuous underlying process and include latent growth dynamic factor analysis and differential equation models. These models are appropriate for data that are continuous or are collected over many measurement occasions.

In addition, point processes, such as the Hawkes process (Halpin & De Boeck, 2013; Hawkes, 1971; Hawkes & Oakes, 1974) also can be used to address issues as in 2.b. Dynamic factor analysis (DFA) is appropriate for modeling dyadic interactions, especially when multivariate time series are available (Ferrer & Nesselroade, 2003; Molenaar, 1985). It combines factor analysis with time series such that it accounts for the autocorrelations

among the multiple variables over time (Browne & Nesselroade, 2005; Ferrer & Zhang, 2009). DFA focuses on time-related influences between the two individuals, where one could identify possible influences from one person's action at a given time to the other person's action at the next occasion. These techniques are particularly useful when the data show fluctuations, without trends or other forms of nonstationarity. Differential equation models (DEMs) have been used to model interactions in dyads (Felmlee, 2006) such as turn-taking in conversations (Buder, 1991; Newton, 1993) and the development of various types of social relationships.

Another versatile class of stochastic models includes Markov models. With a hidden Markov model (HMM) we can model the way the rate of change for each subject also changes over time as different temporal variables affect the way in which subjects respond to stimuli. The HMM can be applied to investigate the changes within individuals over time in VAs.

In the application of HMM to CPS data, observations might be classifications of different student problem-solving strategies with state transitions describing the likelihoods of transitioning from one general problem-solving strategy to another. For example, in a virtual CPS environment, observations might be sequences of online chat between students, and state transitions might describe the communicative roles of students (e.g., facilitator, critic, peer tutor) or the effectiveness of the information sharing and knowledge construction (see Soller & Stevens, 2008).

Yet another new approach to analyze CPS data is to use the Hawkes process to identify the type of events observed during the CPS task, based on the time structure of the interactions. The model can estimate whether an event is spontaneous, auto-dependent, or cross-dependent (Halpin & von Davier, 2013; von Davier & Halpin, 2013). In order to measure the performance of individuals in teams, where the teams change over time in some fashion, advanced modeling techniques from the social network analysis may be considered.

Logistical and Operational Concerns

The validity and utility of VAs will in large part be determined, informed, and improved by the results of actually administering them as well as capturing and analyzing the data they produce. VAs have many new components which, although they can be powerful assessment tools, contain many "moving parts" that must work together in order to be effective. This may require many traditional assessment-developing teams and individuals to rethink how they actually create, administer and analyze assessments. In

well-crafted traditional assessments, assessment developers create items to measure performance along particular constructs. The assessments themselves are then created from an appropriate combination of those items and administered to the appropriate individuals. When the results of the assessments are returned, psychometricians analyze and interpret them with the aid of psychometric and other statistical models.

With VAs, in addition to whether or not a test taker can correctly answer a question, we are interested in the capability of understanding the processes through which test takers arrive at their given answers. As a result, information on things like answer changes and time delays between actions must be recorded and analyzed as well as interactions with the user-interface (UI) with regards to tools built into the task (i.e., dynamic graphs, review panels, audio/video buttons). Serious thought for how those data should and can be used must be taken.

In traditional assessment, individual items and related content are, in many respects, discrete units that are almost entirely what the test taker sees. Furthermore, data capture in traditional assessment is relatively simple, being largely limited to capturing a test taker's final responses. As a result, changing, adding, or removing items and content at many points during the development process is accomplished relatively easily, adds comparatively little cost in time or money, and requires little if any change to data capture considerations. Not surprisingly, some clients of assessment developing organizations have enjoyed the flexibility that traditional assessments afford for late changes to assessment content. Such "last-minute" changes to assessments can be very costly in VAs, however.

With VAs, changes to items and content have the added costs of potentially having to restructure the data capture, scoring rubrics, and scoring procedures to accommodate new data types. The degree of difficulty/cost can range depending on the type of item/content changes. The measurable and measured importance of test-taker interactions through the task and the vulnerability of some of the user interface (UI) and data-capture aspects of each assessment to flaws unintentionally introduced with each change necessitates that the entire task undergo a quality assurance (QA)/quality control (QC) process following any changes. This also can multiply the cost for making changes at many points during the development of VAs.

In order to keep costs under control, both assessment developing organizations and their clients have to rethink their behavior, particularly regarding how and when to make changes to assessments during the development process. They must keep in mind that VAs measure new things using new technology or new applications of older technology and, as a result, must be designed from the beginning in new ways.

The (Right) Time for Introducing Changes in an Assessment

Figure 1.4, offers a simplified picture of the development of VAs. Not every assessment production process will contain the same number of actors, but they should all contain the functions of those actors. In traditional assessments, a client needs an assessment designed, produced, administered, analyzed, and its results reported. For our purposes we will call the actor responsible for these tasks the assessment developer as in Figure 1.4. The process should be iterative and the client typically approves of or requires changes to the assessment at various times, restarting the cycle from various points. Ideally, much of this iteration occurs prior to the vendor delivering a polished product.

The chief differences between VA development and traditional assessment development emanate from the interrelatedness of items in VAs and the need for the developer also to consider what additional data to capture, how to capture it and how best to analyze all of those new data for the user's interaction with the assessment. The effect of all of this is that changes, like adding a new item or significantly changing an item, can greatly impact the task and even reset the development cycle. A further complication is that rarely does any single actor possess the necessary resources to perform all of the tasks necessary to develop a VA or to do so at sufficient scale. Thus, other actors (subcontractors, vendors, etc.) are often delegated portions of the assessment development tasks (Figure 1.4).

The timing of changes to assessments is crucial since reiterations originating at various stages of an assessment's development can have very different costs and risks. For instance, in scenario and game-based assessments (SGBAs) it is extremely important that later changes do not disrupt the narrative already in place. A disruption of that sort could seriously set the developer back in the development cycle. In the most extreme cases, it could effectively restart the entire cycle resulting in major resource losses for the all parties involved. There are similar concerns for other types of VA. Thus early stage cross-functional participation and revision is essential here.

The principle of properly timing changes to VAs is not very different from the logic dictating when changes should be made to traditional assessments: changes should be made when necessary and where they would be least costly. In practice, ensuring that changes are made at an appropriate time can be quite complex. Each actor involved has its own internal calendar according to which it will act unless a common calendar is established; those separate calendars may not be fully understood by all parties involved. The current innovative nature of VA, however, makes adhering strictly to any such common calendar difficult: assessment developers may find that innovative features of a VA, which may have seemed simple

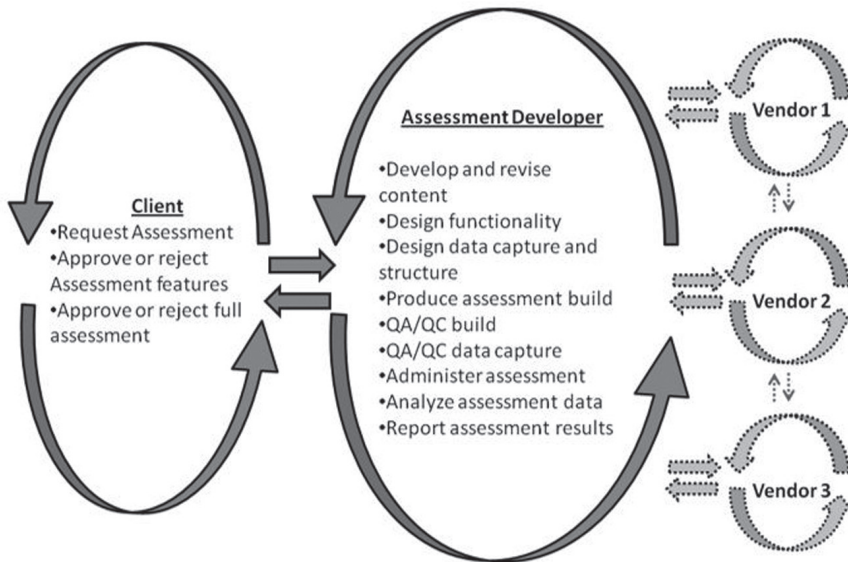


Figure 1.4. Simplified illustration of actors involved in developing a virtual assessment and their respective duties. The duties in the second block may be assumed by a single assessment developer, but certain duties or portions thereof may be sub-contracted to vendors. The circular arrows indicate that the development process iterates and that the number of iterations may vary for different actors. Arrows with dotted borders indicate relationships and iterations which may not necessarily exist. It is worth noting that in this simplified figure, many of the tasks assigned to a given party represent iterative process with their own iterative subprocesses, for example, developing and revising content subsumes the distinct processes of developing the content scenario/storyline/etc., developing the user experience, developing the content basis of the assessment, and ensuring strong connections of the task to evidence models and reporting goals.

to deliver at the outset, have unanticipated complications which require more time to address. Both of these factors can cause delays disrupting calendar alignments and complicating finding appropriate times to implement changes. In order to address the coordination and timing issues that this can introduce, the common calendar developed would need to be very flexible. We can, however, mitigate the likelihood of having to change the calendar drastically by having input from appropriate experts “in the room” to determine how much time will actually be needed at various stages of development.

Overall, cross-functional participation in the early stages of development reduces the risk of and need for later stage changes. Utilizing the ECD model does not solve the problem of these dependencies, but it can

help to make them explicit. This, in turn, helps the actors involved in the development process to understand better the costs of changes at various points in the development cycle, giving them the power to make more informed decisions about making changes.

EXAMPLES OF VIRTUAL ASSESSMENTS

In order to anchor the discussion about VAs, we include here a brief description of three examples of assessments: the Nation Assessment of Educational Progress (NAEP) Technology and Engineering Literacy (TEL) assessment, which is a scenario based assessment, the GlassLab's first game-based assessment, SimCity-Edu: Pollution Challenge!, and the Educational Testing Service (ETS) Collaborative Science Assessment, the Tetralogue. The descriptions are brief but we provide links to published work for the interested reader. In addition to describing the assessments we also illustrate different aspects of the VAs that were discussed in the previous sections, such as the role of ECD in the development of VAs, the log files, and the special analyses conducted on the process data. The three examples are also ordered with respect to their complexity and data dependencies, from scenario-based item types to game-based assessment, and virtual collaborative assessments.

NAEP TEL's Wells Task²

The Wells task is an interactive, scenario-based task developed for eighth-grade participants in the Nation Assessment of Educational Progress (NAEP) Technology and Engineering Literacy (TEL) assessment. The task itself can be found at http://nces.ed.gov/nationsreportcard/tel/wells_item.aspx. NAEP TEL aims to assess student content knowledge and cross-cutting practices. The NAEP TEL content areas of Design and Systems, Technology and Society, and Information and Communications Technology are cross-cut by practices labeled Understanding Technological Principles, Communicating and Collaborating, and Designing Solutions and Achieving Goals. To assess these practices, the NAEP alliance (Educational Testing Service, Business Intelligence, Inc, Fulcrum IT Services Company, and Pearson) created several VAs and computer-based traditional assessments, with survey questionnaires, which were administered to a representative sample of U.S. students. The Wells task is the first of these pilot VAs to be released to the public (see Figure 1.5).

The Well's task focused on the TEL Design and Systems competencies (www.nagb.org) using a scenario in which students are asked to help repair a well-pump. Students are given an overview of how well-pumps and the aquifers from which they draw work. They are then asked to use their

knowledge and skills to address one that is malfunctioning in a remote Nepalese village. During this highly scaffolded and linear scenario, students diagnose, troubleshoot, repair, and identify a maintenance plan for the pump.



Figure 1.5. Pump Repair screen from the NAEP TEL Wells task.

The Wells task included both traditional items, in which students were asked to respond to explicitly presented questions, and VA items, in which assessment was more discrete. The Wells task included digital tools and resources with which students could interact to solve problems. The process data generated from those nonscored interactions between the students and the virtual environment provided the basis for making inferences about student's final answers and how they decided upon them. In certain sections of the Well's task the interactions themselves were scored as outcome data, since the behavior patterns demonstrating better or poorer understanding could be defined in the scoring rubric—that is, efficiency and systematicity scores in the pump repair section of the task (Figure 1.5. (Keehner et al., 2014).

In order to analyze the process/sequence data collected from this interactive task, different types of analyses were conducted. Bergner, Shu, and von Davier (2014) explored visualization and clustering techniques with respect to sequence data from the Wells task. Visualization issues included representing progress towards a goal and accounting for variable-length sequences. Clustering issues focused on external criteria with respect to official scoring rubrics of the same sequence data (i.e., efficiency and sys-

tematicity scores mentioned earlier). The goal was to understand to what extent clustering solutions align with score categories. They found that choices related to data preprocessing, distance metrics, and external cluster validity measures all impact agreement between cluster assignments and scores.

Students' activities can be characterized by a sequence of time-stamped actions of different types with different attributes. For a task in which only the order of the actions are of great interest, the process data can be well characterized as a string of characters (action string, hereafter) if we encode each action name as a single character. In a different paper, Hao, Shu, and von Davier (2015) reported on evaluating students' performances by comparing how far their sequences of action strings are from the action sequence that corresponds to the best performance, in which the proximity is quantified by the edit distance between the strings (a measure inspired by the work conducted in text analysis and natural language processing). The Levenshtein distance, which is defined as the minimum number of insertions, deletions and replacements needed to convert one character string to another, was used in this study. The results showed a strong correlation between the edit distances and the scores obtained from the scoring rubrics, implying the edit distance to the best performance sequence can be considered as a new feature variable that encodes information about students' proficiency, and shedding light on the value of data-driven scoring rules for test and task development as well as for refining the scoring rubrics.

SimCity-Edu: Pollution Challenge!³

GlassLab is a research and development collaborative effort of the Institute of Play, Electronic Arts, the Entertainment Software Association (ESA), Educational Testing Service (ETS), and Pearson, funded by the Bill and Melinda Gates and the John D. and Catherine T. MacArthur Foundations, to create game-based assessments. SimCity-Edu: Pollution Challenge! was the first product of that collaboration (see Figure 1.6).

Pollution Challenge is a modified version of SimCity, a simulation game by Maxis that lets players plan, build, and "run" digital cities populated with digital agents that mirror the decisions and activity of their real-life counterparts. Pollution Challenge presents students with opportunities to build and create based on an understanding of systems and human impact on the environment. Unlike the open "sandbox" approach in the commercial games, students do not have to build the cities in Pollution Challenge. Instead, the missions on which students embark include constrained, pre-designed cities each with a specific problem that students must remedy. In one mission, students must reduce the level of pollution and increase the



Figure 1.6. SimCity-Edu: Pollution Challenge! Task.

level of employment. Successful players are those who can employ green energy technologies or rezone areas (e.g., shifting industrial to commercial) to reduce pollution in each city while simultaneously supporting that city's job growth. Through these missions, students are introduced to themes of human impact on the environment as presented in the Next Generation Science Standards (NGSS Lead States, 2013) core disciplinary ideas and the NGSS's cross-cutting concept of systems and systems models.

Collaborative Science Assessment, the Tetralogue

The Tetralogue, a Collaborative Science Assessment, developed at Educational Testing Service (see Figure 1.7), embedded a rich simulated collaborative problem-solving (CPS) task in a science assessment along with a traditional science test that consists of multiple choice (MC) items, a background questionnaire (BQ), a personality test, and a postadministration survey to gauge quality of the experience with the virtual collaborative task (Hao, Liu, von Davier, & Kyllonen, 2015; Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2016). The Tetralogue was based on the Science Trialogue task (Zapata-Rivera et al., 2014). It was developed according to ECD principles and had a twofold purpose: to jointly assess the science skills through CPS tasks and MC items, and to assess the CPS skills themselves through the CPS tasks. The expectation was that the two dimensions would correlate, but not be identical. The CPS task is a simulation of a science lesson about volcanoes and requires the collaboration of two human students in solving a problem in a virtual environment via a chat box. The data from the interactions of the two humans are rich time series and can

be modeled statistically using appropriate methods as described in von Davier and Halpin (2013).

In the collaborative setting of the Tetralogue, dyads of students work collectively to make a prediction of a volcano alert level. Within the collaborative task, a structured system prompt was designed to facilitate the collaborative discourse between dyad participants. The data were collected in a log file designed especially to allow for the appropriate psychometric analyses. The pilot was conducted using Amazon's Mechanical Turk (www.mturk.com) with a selected sample of 500 dyads (see Liu et al., 2016 for details).

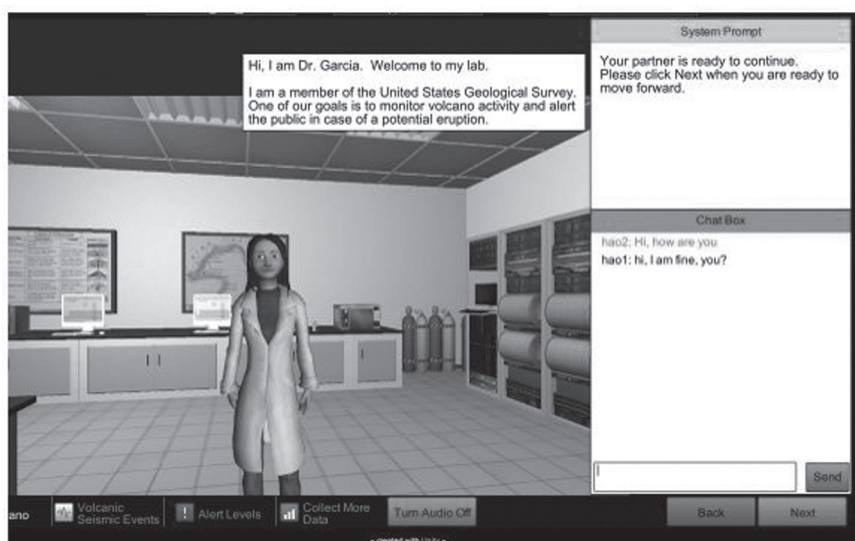


Figure 1.7. The Tetralogue: A Collaborative Science Assessment developed at Educational Testing Service

DISCUSSION

In this chapter we briefly reviewed the advantages and disadvantages of VAs, and indicated some of the considerations the assessment experts need to make at different stages in the life of the assessment. We touched only briefly on other important aspects of the VAs: reliability, generalizability, validity, and comparability of tasks over time. There is still significant research needed to properly address these aspects of VAs. For instance, typical approaches to traditional psychometric analysis do not take into account a test taker's cognition, an issue which may begin to be addressed

with tools such as cognitive task analysis or computational modeling (Williamson et al., 2004). Many of the psychometric aspects were discussed in detail in the framework of game-based assessments and simulation-based assessments in Mislevy et al. (2014) and von Davier and Mislevy (2015). However, new types of methods borrowed from the artificial intelligence and machine learning are showing promise when merged with the traditional psychometrics. von Davier (2015) introduced the term “computational psychometrics” to encompass the merger of these methodologies for analyzing process data using both data-driven and theory-driven approaches.

We end this chapter by mentioning that the advantages of the VAs and the rigor of the psychometric requirements are more or less salient depending of the use of the assessment: for an assessment that is meant to be formative and for which the stakes are low, some of these psychometric requirements may hold to a lesser degree and the assessment may still be useful for its purpose.

Similarly, one may consider the issues of test fairness, test security and test takers’ privacy in the context of VAs. These may represent a topic for another work. As with the psychometric requirements, it is worth noting that these issues may be less problematic for some test uses than for others.

ACKNOWLEDGMENT

This chapter is based on the invited presentation of Alina A. von Davier at MARCES conference 2014 (von Davier, 2014). The opinions presented in this chapter are those of the authors and not necessarily of Educational Testing Service. The authors thank Deirdre Kerr, Jiangang Hao, Zhang Zhu, Greg Vafis, and Shu-Kang Chen for their suggestions on the previous version of the manuscript. Correspondence should be sent to Alina A. von Davier at avondavier@ets.org

NOTES

1. There are some new approaches to this type of interaction, but there is not a rich literature on this topic and there are almost no applications in educational assessment.
2. For more information on the Wells task, TEL Framework, or NAEP, we direct the interested reader to the following resources:
http://nces.ed.gov/nationsreportcard/pdf/about/schools/Grade8_TEL_FactSheet.pdf
<http://www.nagb.org/content/nagb/assets/documents/publications/>

frameworks/technology/2014-technology-framework-abridged.pdf
www.nces.ed.gov/nationsreportcard

3. For more information on the ECD and psychometrics for this game, see Mislevy et al. (2014) and <http://www.glasslabgames.org/games/SC>. For more information on the data management and log file see Hao, Mislevy, von Davier, and Smith (2016).

REFERENCES

- Almond, R., Steinberg, L., & Mislevy, R. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *The Journal of Technology, Learning and Assessment*, 1(5), 1–64.
- Bergner, Y., Andrews, J. J., Zhu, M., & Kitchen, C. (2015, July). *Agent-based modeling of collaborative problem solving*. Paper presented at the 10th Annual INGroup Conference, Pittsburgh, PA.
- Bergner, Y., Shu, Z., & von Davier, A. A. (2014). *Visualization and confirmatory clustering of sequence data from a simulation-based assessment task*. Proceedings of the Educational Data Mining Conference, London, England. Retrieved from <http://www.educationaldatamining.org/conferences/index.php/EDM/2014/paper/view/1366/0>
- Blech, C., & Funke, J. (2005). *Dynamis review: An overview about applications of the Dynamis approach in cognitive psychology*. Bonn, Germany: German Institute for Adult Education.
- Blech, C., & Funke, J. (2010). You cannot have your cake and eat it, too: How induced goal conflicts affect complex problem solving. *The Open Psychology Journal*, 3, 42–53.
- Browne, M. W., & Nesselroade, J. R. (2005). Representing psychological processes with dynamic factor models: Some promising uses and extensions of ARMA time series models. *Psychometrics: A festschrift to Roderick P McDonald*, 415–452.
- Buder, E. H. (1991). A nonlinear dynamic model of social interaction. *Communication Research*, 18, 174–198.
- Chung, G. K. W. K., Baker, E. L., Vendlinski, T. P., Buschang, R. E., Delacruz, G. C., Michiuye, J. K., & Bittick, S. J. (2010, April). Testing instructional design variations in a prototype math game. In R. Atkinson (Chair), *Current perspectives from three national R&D centers focused on game-based learning: Issues in learning, instruction, assessment, and game design*. Structured poster session at the meeting of the American Educational Research Association, Denver, CO.
- Felmlee, D. H. (2006). Interaction in social networks. In J. De Lamater & A. Ward (Eds.), *Handbook of social psychology* (pp. 389–409). New York, NY: Springer.
- Ferrer, E., & Nesselroade, J. R. (2003). Modeling affective processes in dyadic relations via dynamic factor analysis. *Emotion*, 3, 344.
- Ferrer, E., & Zhang, G. (2009). Time series models for examining psychological processes: Applications and new developments. In R. Millsap, & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 637–658). London, England: SAGE.

- Gibbons, R. D., & Hedeker, D. R. (1992). Full-Information Item Bi-Factor Analysis. *Psychometrika*, 57, 423–436.
- GlassLab Games. (2015). Retrieved from <http://www.glasslabgames.org/games/SC>
- Halpin, P. F., & De Boeck, P. (2013). Modeling dyadic interaction with Hawkes process. *Psychometrika*, 78, 793–814.
- Halpin, P. F., & von Davier, A. A. (2013, April). *Evaluating the roles of individual members in team interactions*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.
- Hao, J., Liu, L., von Davier, A. A., & Kyllonen, P. (2015) Assessing problem solving with simulation based tasks. *International Journal of Computer-Supported Collaborative Learning*.
- Hao, J., Shu, Z., & von Davier, A. A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining*, 7, 33–50. Retrieved from <http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/JEDM072>
- Hao, J., Smith, L., Mislevy, R., Davier, A., & Bauer, M. (2016). Taming log files from game/simulation-based assessments: Data models and data analysis tools. *ETS Research Report Series*, 2016(1), 1–17.
- Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58, 83–90.
- Hawkes, A. G., & Oakes, D. (1974). A cluster representation of a self-exciting process. *Journal of Applied Probability*, 11, 493–503.
- Katz, I. R., Martinez, M. E., Sheehan, K. M., & Tatsuoaka, K. K. (1998). Extending the rule space methodology to a semantically-rich domain: Diagnostic assessment in architecture. *Journal of Educational and Behavioral Statistics*, 23, 254–278.
- Keehner, M., Agard, C., Berger, M., Bertling, J., & Shu, Z. (2014). *Analyzing interactivity, performance, and background data from the National Assessment of Educational Progress (NAEP) Technology & Engineering Literacy (TEL) Wells Task*. Federal Research Memorandum on NAEP Task Component, U.S. Dept. of Education, Contract Award No. ED-IES-13-C-0015.
- Klieme, E. (2004). Assessment of cross-curricular problem-solving competencies. In J. Moskowitz & M. Stephens (Eds), *Comparing learning outcomes. International assessments and education policy* (pp. 81–107). London, England: Routledge.
- Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, J.-D. (2016). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on computational tools for real-world skill development*. Hershey, PA: IGI-Global.
- McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., & Scalese, R. J. (2010). A critical review of simulation-based medical education research: 2003–2009. *Medical Education*, 44, 50–63.
- Mislevy, R. J. (2011). *Evidence-centered design for simulation-based assessment*. (CRESST Report 800). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, University of California at Los Angeles.
- Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., ... John, M. (2014). *Psychometric considerations in game-based assessment*. Redwood City, CA: GlassLab.

- Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. Lane, M. Raymond, T. M. Haladyna, & S. M. Downing (Eds), *Handbook of test development* (pp. 61–90). London, England: Routledge.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–62.
- Molenaar, P. C. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, 50, 181–202.
- Newton, D. (1993). A dynamic systems approach to development: Applications. In L. B. Smith & E. Thelen (Eds), *MIT Press/Bradford Books series in cognitive psychology* (pp. 241–264). Cambridge, MA: MIT Press.
- NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: National Academies Press.
- Polvichai, J., & Khosla, P. (2002). An evolutionary behavior programming system with dynamic networks for mobile robots in dynamic environments. In *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and System*, 1, 978–983.
- Press, G. (2013). A very short history of data science. *Forbes.com*. Retrieved July 28, 2013, from <http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/>
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Sokolowski, J. A., & Banks, C. M. (Eds.). (2009). *Principles of modeling and simulation: A multidisciplinary approach*. Hoboken, NJ: Wiley.
- Soller, A., & Stevens, R. (2008). Applications of stochastic analyses for collaborative learning and cognitive assessment. In G. R. Hancock & K. M. Samuelsen (Eds), *Advances in latent variable mixture models* (pp. 217–253). Charlotte, NC: Information Age Publishing.
- von Davier, A. A. (2014, October). *The virtual world and the reality of testing: Simulation- & game-based assessments*. Paper presented at the Maryland Assessment Conference, College Park, MD.
- von Davier, A. A. (2015, July). *Virtual & collaborative assessments: Examples, implications, and challenges for educational measurement*. Invited Talk at the Workshop on Machine Learning for Education, International Conference of Machine Learning, Lille, France http://dsp.rice.edu/ML4Ed_ICML2015
- von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations*. (Research Report 13-41). Princeton, NJ: Educational Testing Service.
- von Davier, A. A., & Mislevy, R. (2015). Design and modeling framework for the 21st century simulations- and game-based assessments. In C. Wells & M. Faulkner-Bond (Eds.) *Educational measurement: From foundations to future*. New York, NY: Guilford.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data*. (Research Report 5-16). Princeton, NJ: Educational Testing Service. Retrieved from: <http://www.ets.org/Media/Research/pdf/RR-05-16.pdf>
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

- Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., Behrens, J. T., & DeMark, S. F. (2004). Design rationale for a complex performance assessment. *International Journal of Testing*, 4, 303–332.
- Yao, L. (2008). BMIRT: Bayesian Multivariate Item Response Theory (Version 1.0) [Computer Software]. Monterey, CA: CTB/McGraw-Hill.
- Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., ... Yukhymenko, M. (2012). Our princess is in another castle a review of trends in serious gaming for education. *Review of Educational Research*, 82, 61–89.
- Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., & Katz, I. R. (2014). Science inquiry skills using dialogues. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International conference on Intelligence Tutoring Systems* (pp. 625–626). New York, NY: Springer.

