Journal of Computerized Adaptive Testing

Volume 10 Number 1

February 2023

An Extended Taxonomy of Variants of Computerized Adaptive Testing

Roy Levy, John T. Behrens, and Robert J. Mislevy

DOI 10.7333/2302-100101

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

[**www.iacat.org**](http://www.iacat.org)

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing
Volume 10, Number 1, February 2023
DOI 10.7333/2302-100101
ISSN 2165-6592

An Extended Taxonomy of Variants of Computerized Adaptive Testing

Roy Levy
University of Maryland

John T. Behrens
Cisco Systems

Robert J. Mislevy
University of Maryland

Preface

Robert J. Mislevy
University of Maryland (Emeritus)

It was my privilege to present a keynote address to the 2014 Computerized Adaptive Testing Summit, hosted in Princeton, New Jersey, by the International Association for Computerized Adaptive Testing (IACAT). In the address, I described a taxonomy of adaptive testing schemes that Roy Levy, John Behrens, and I wrote about in greater detail in “Variations in Adaptive Testing and Their Online Leverage Points,” which was published as chapter XI in David Williams, Mary Hricko, and Scott Howell’s (2006) edited volume *Online Assessment, Measurement and Evaluation: Emerging Practices*. With the kind permission of IGI Global, the publisher, that chapter is reprinted here.

The aim of the presentation was to place key ideas of computerized adaptive testing (CAT) as applied in educational and psychological measurement—namely, adaptivity, probabilistic reasoning, and knowledge-based model construction—into a broader conceptual framework. The framework draws on David Schum’s (1994) principles of evidentiary reasoning and Glenn Shafer’s (1976) notion of frames of discernment in statistics. Traditional every-student paper-and-pencil tests and item response theory (IRT)-based CAT are seen as exemplars in a structured space of potential adaptive assessment paradigms, with organizing concepts that help us better understand not only these familiar forms of assessment but a range of others that, despite wide-ranging surface differences, build on the same underlying principles, especially as we see supporting technologies continue to emerge. I believe that continuing developments in assessment, building on advances in computation, analytic methods, and psychometric models, and forms such as simulation-based tasks, game-based assessment, and collaborative assessment, benefit from the perspectives offered here (for examples of these developments, see Baker et al., 2017; Ifenthaler & Kim, 2019; Ke et al., 2019; von Davier et al., 2017, 2021; and Yan et al., 2020).

Abstract

This paper builds on foundational work on probabilistic frames of reference and principled assessment design to explore the role of adaptation in assessment. Assessments are characterized in terms of their claim status, observation status, and locus of control. The relevant claims and observations constitute a frame of discernment for the assessment. Adaptation occurs when the frame is permitted to evolve with respect to the claims or observations (or both); adaptive features may be controlled by the examiner or the examinee. In describing the various combinations of these characteristics, it is argued that an online format is preeminent for supporting common and emerging assessment practices in light of adaptation.

Keywords: adaptive testing, evidence-centered design, item response theory, knowledge-based model construction, missingness

The digital revolution has brought dramatic shifts in the activities and conceptualizations of modern life by providing easily transformable digital representation, dramatic computing power for calculation and decision making, and the use of large-scale databases. Internetworking technologies are having a similarly impressive impact by allowing geographically and computationally distributed combinations of this representational, computing, and database power. Having inherited many of our current conceptualizations and tools from predigital and pre-networked times, it is prudent to re-examine our understandings and language in light of these new possibilities. In the context of a globally linked digital world of computation, representation, and data, one area of potentially great benefit is that of computerized-adaptive testing. Online presentation is greatly affected by the simulation and display technologies that continue to emerge; task selection is greatly affected by the availability of computing power and the availability of databases that may need to be remote from the user.

As the assessment community moves forward in harnessing these opportunities, it is important that discussion not only occur in the language and dimensions inherited from a predigital era, but that we re-examine the language and categories available to us to take advantage of the wide range of possibilities at hand. The focus of this work is to lay out a conceptual framework and taxonomy of adaptive assessment based on discussion of probabilistic frames of reference (Shafer, 1976) and dimensions of evidentiary reasoning that serves as the foundation for modern assessment (Mislevy et al., 2003). This will be addressed in the context of a pragmatic delivery model (Almond et al., 2002) that has been embedded in industry computing standards and in large-scale online assessment systems (Behrens et al., 2004, 2006).

There is no shortage of ways to classify assessments. One may consider assessments in terms of classical test theory (CTT) versus IRT, linear versus adaptive, large scale versus small scale, high stakes versus low stakes, diagnostic/formative versus summative, and of course, computer-based versus paper and pencil. Building from Shafer's (1976) conception of a "frame of discernment" in probability-based inference and Mislevy et al.'s (2003) work on "evidence-centered" assessment design, we propose a taxonomy that differentiates assessments along the three dimensions of (a) *observation status*, (b) *claim status*, and (c) *locus of control*. This foundation allows us to highlight the inferential roles that adaptivity can play in assessment. It offers a principled perspective for examining advantageous features of various adaptive testing models such as reduced time and increased precision in adaptive observation assessments, and diagnostic capability in examinee-controlled assessments. In detailing the taxonomy, we point out ways in which online assessment enables or enhances these features.

Conceptual Foundations

Frames of Discernment

In his 1976 treatise, *A Mathematical Theory of Evidence*, Glenn Shafer defined a frame of discernment as all of the possible subsets of combinations of values that the variables in an inferential problem at a given point in time might take. The term "frame" emphasizes how a frame of discernment effectively delimits the universe in which inference will take place. As we shall see in the next section, the frame of discernment in assessment comprises student-model variables and observable variables. The former concern aspects of students' proficiencies such as knowledge, skill, ability, strategies, behavioral tendencies, and so on; the latter concern aspects of things they say, do, or make that provide clues about their proficiencies. The term "discernment" emphasizes how a frame of discernment reflects purposive choices about what is important to recognize in the inferential situation, how to categorize observations, and from what perspective and at what level of detail variables should be defined.

A frame of discernment depends on beliefs, knowledge, and aims. Importantly, in everyday inferential problems as well as scientific problems, frames of discernment evolve, as beliefs, knowledge, and aims unfold over time. The need for vegetable dip for a party may begin with the inferential problem of whether or not there is any on hand and evolve to determining where to go to make a purchase and which brand and size to buy. People move from one frame of discernment to another by ascertaining the values of some variables and dropping others, adding new variables or refining distinctions of values of current ones, or constructing a rather different frame when observations cause them to rethink their assumptions or their goals.

Evidence-Centered Assessment Design

Evolving complexities in many aspects of assessment, from development and administration to scoring and inferences, have rendered many popular terms as ambiguous or limited in scope at best and irrelevant at worst. Expressions such as "item," "answer," and "score," which on the surface appear to be quite general, are, in fact, limited in their application and do not suffice for describing or designing complex, innovative assessments (Behrens et al., 2004). Assessments that are innovative, either in terms of how they are developed and administered and/or what they demand of the examinee, are better served by a richer, more general terminology. The language of evidence-centered design (ECD; Mislevy et al., 2003), a descriptive and prescriptive framework for assessment development and implementation, provides such a terminology. A full explication of ECD and its components is beyond the scope and intent of this paper. This section provides a brief description sufficient to introduce terms relevant to the taxonomy.

Evidentiary and Inferential Language

A *claim* is a declarative statement about what an examinee knows or can do. Claims may be broad (the examinee can subtract) or specific (the examinee can subtract negative improper fractions). The level of specificity of the claim(s) is often tied to the purpose of the assessment. Formative assessments often refer to highly specific claims, while summative assessments tend to have broader claims. Claims are hypotheses about the examinee; addressing these hypotheses is the goal of the assessment. Information pertinent to addressing the claims is accumulated in terms of student-model variables, which are typically latent variables representing the underlying construct of interest. The student-model variables represent examinees' knowledge, skills, or abilities and are therefore the targets of inference in an assessment.

In order to gain information regarding the student-model variables and the claims to which they are relevant, *observations* must be collected to serve as evidence. The conceptual assessment framework addresses the components of the psychometric models; the components are assembled to manage the collection and synthesis of evidence in an operational assessment, including student-model variables and observable variables. It includes an assembly model that contains the logic used to choose tasks to provide observations.

A Language and Model for Assessment Delivery

The four-process delivery system (Almond et al., 2002) provides a description of the interaction among processes for delivering tasks, characterizing performance, updating belief about examinees, and, when appropriate, selecting subsequent tasks in light of what has been learned thus far.

Task selection is the first of these four processes and consists of activities designed to determine which task should be presented to the examinee. In simple cases, this might be a non-deterministic rule such as “show the next question” or may consist of a complex selection algorithm optimizing (or minimizing) values to make choices along a number of dimensions.

Task selection leads into the **presentation process,** which consists of the presentation of task materials to the examinee. The result of the task/examinee interaction results in a record that is called a work product. Many assessment practitioners refer to this result as an answer; however, answer implies a question and in many situations, there is no question, but rather a task to be completed. Consider, for instance, an examinee that is asked to draw a painting or perform a dance. In this case, there is presentation (the instructions and the stage), and the dance or painting is the work product.

The third process in the model consists of examining the **work product** and characterizing the work product in terms of one or several variables or observables. These observables are the computing instantiation of the observations desired from an inferential perspective. This transformational process is called either evidence identification or response processing. Here again, the language is extremely flexible. A more traditional language might talk about scoring a question. This is one way to look for characteristics of the work product (the correct answer) but is overly restrictive. The four-process terminology allows what is being looked for in the work product to be any desired characteristic. Perhaps efficiency or the presence of a particular strategy are relevant features of a work product.

Evidence accumulation is the fourth process. This refers to statistical processes (however simple or complex) that are used to synthesize or accumulate data from the observables to estimate values for the student-model variables that inform our knowledge of the claims we want to make.

Each of these processes requires data and a database to keep the data. Information about task difficulty and task domain is generally needed in task selection. Increasingly complex presentation (video, simulation) requires increasingly large databases, and scoring and statistical computing require unique computing resources as well. The four-process model refers to a “task/composite” library as a location for such information.

The four-process model is especially relevant to assessment computing in the current online age because its language allows for great complexity throughout the assessment delivery process. In this scheme, the unit of analysis is observables of any value, rather than integer values associated with questions. Accordingly, complex tasks can lead to multiple observables, which may load on multiple student-model variables. As the complexity of computer-based assessment expands, the language of the four-process model has already provided a framework to describe it.

At an architectural level, the four-process model suggests that each of the different processes can occur in different locations. For example, in the early versions of the NetPASS, a computer network troubleshooting assessment (Behrens et al., 2004), task selection was made by students in any web browser throughout the world, while presentation came from a remote computer network in the eastern United States, and evidence identification and evidence accumulation happened in a data center in the western United States. As illustrated in this application, appropriately delivered online assessment has the possibility of bringing the best computational, representational, and data resources to bear on the assessment task as needed.

Integration of Epistemic, Inferential, and Delivery Languages

Combining these multiple languages allows us to reframe the discussion regarding adaptive testing at a number of levels. In this scheme, the student-model variables and observable variables in play at a given point in time entail a frame of discernment, which is here characterized as being fixed or adaptive. Here, *fixed* conveys that a particular aspect (i.e., claims or observations) of the frame of discernment is set a priori and is not subject to change. *Adaptive* means that the frame of discernment may evolve in response to unfolding information as assessment proceeds. A fixed-claim adaptive-observation assessment is one in which the claim(s) to be investigated is (are) set in advance and not subject to change, but the set of observations that will be collected to bring to bear on a claim may change during the assessment. In an adaptive-claim assessment, the inferential goals or targets are subject to change; the hypotheses of interest that are investigated may change as new information (from observations) is incorporated.

An important aspect of what are characterized as adaptive assessments is the role of feedback in the task selection or claim selection mechanisms. Feedback into the assessment system (usually to the examiner, possibly also to the examinee) may serve as an aid or as guidance as to how the assessment evolves. For example, it is common practice in computerized adaptive testing to use correctness of response to aid in choosing the next task on the basis of maximizing information regarding a student-model variable. While the majority of adaptive assessments discussed in this paper involve feedback, there are assessments that are characterized as adaptive that do not involve feedback. These points are further elaborated in a later section in the context of an example of one such assessment.

Fixed tests, varieties of existing adaptive tests, and other configurations not in current use can all be described in these terms. Our main interest will be in the areas of task selection and the presentation process, though along the way aspects of evidence identification and evidence accumulation for updating beliefs about examinees will become indispensable.

The Taxonomy

We propose a taxonomy of assessment that classifies assessments based on (a) whether the claims are fixed or adaptive, (b) whether the observations are fixed or adaptive, (c) the control of the claims, and (d) the control of the observations. As will be seen, these dimensions can be understood in terms of how the frame of discernment evolves (or does not) over the course of an assessment. Traditionally, the examiner (or a proxy thereof) has control of the assessment, specifically, what claims to make and what observations to collect. In what follows, we discuss these familiar situations and departures from them. The taxonomy is represented in Figure 1 (where short descriptors of examples are given for the most interesting cells). The 16 cells represent the possible combinations of claim status (fixed vs. adaptive and examiner- vs. examinee-controlled for both cases) and observation status (fixed vs. adaptive and examiner- vs. examinee-controlled for both cases).

Figure 1. All Possible Combinations of the Taxonomy

			Observation Status			
			Fixed		**Adaptive**	
			Examiner	**Examinee**	**Examiner**	**Examinee**
Claim status	**Fixed**	**Examiner**	1. Usual, linear test	2	3. CAT	4. SAT
		Examinee	5	6	7	8
	Adaptive	**Examiner**	9. MMPI–2—Examiner decides how to pursue analysis	10	11. Examiner chooses target, multidimensional CAT	12. Examiner chooses target, multidimensional SAT
		Examinee	13. MMPI–2—Examinee decides how to pursue analysis	14	15. Examinee chooses target, multidimensional CAT	16. Examinee chooses target multidimensional SAT

Note. CAT = computerized adaptive test; MMPI–2 = Minnesota Multiphasic Personality Inventory–2; SAT = self-adaptive test.

The presentation is organized as follows. In proceeding through the various combinations, the order will follow the rows of Figure 1. For each row, common features that apply to all assessment types in the row will be presented. Distinctions between the assessment types in each row will be more specifically addressed in subsections.

Fixed, Examiner-Controlled Claims

The most familiar types of assessment are of this kind. The claims are fixed in the sense that inferences are made regarding the same claims for each examinee; the assessment is developed to arrive at (estimates of) values of the same student-model variables. The claims are examiner controlled in that the examiner, rather than the examinee, determines the targets of inference.

Examples of these types of assessments discussed below include fixed linear assessments, computerized adaptive tests, and self-adaptive tests.

Cell 1. *Fixed, examiner-controlled observations.* Traditional assessments in which the tasks presented to each examinee are determined by the examiner a priori are of this kind. The sequence of tasks may also be determined by the examiner a priori or it may be random, such as in cases where there is concern for test security. Regardless, the observables are fixed in the sense that evidence for the values of the student-model variables come from values of the same observables for all examinees. The observables are examiner controlled in the sense that the examiner, rather than the examinee, determines the tasks, possibly also including the sequence of tasks, that are presented. In Shafer's (1976) terms, the examiner has determined a frame of discernment, encompassing the same fixed set of student-model variables and observable variables for all examinees. Neither the frame of discernment nor the gathering of evidence varies in response to realizations of values of observable variables or their impact on beliefs about student-model variables.

An example of these assessments is a fifth-grade spelling test that asks students to spell the same words in a sequence devised by the teacher such that the teacher can make inferences about the same proficiency for each examinee. Another example is a statewide math assessment where the same set of IRT-scaled tasks are given to all examinees in the same order to obtain estimates of the latent mathematics ability of each student. This classification may include assessments that vary with respect to any number of striking and often important dimensions, such as high stakes versus low stakes, summative versus formative, online versus paper and pencil, and CTT-based versus IRT-based. What's more, this classification subsumes the configuration that is popularly misperceived as encompassing all possible assessments: A set of tasks is developed, given to all examinees, and scored to make inferences/decisions about the same qualities of the students.

Assessments of this type were developed long before the invention of computers. Gains from online administration of this type may be in terms of improved test security; a reduction in coding, scoring, and associated measurement errors; increased data storage; and immediate score reporting (Bunderson et al., 1988). In terms of supporting the assessment argument (i.e., the warranted inference regarding the claim), the use of an online administration does little; indeed, this class of assessments was developed and refined before the advent of online assessment and, therefore, does not involve features of other assessment systems that rely on an online format.

One notable advantage that may be achieved by an online administration in even this class of assessments is the possibility of innovative task types. Tasks that involve moving parts or audio components typically cannot be administered without a computer. They almost certainly cannot be administered in a standardized way absent a computer. To the extent that innovative tasks enhance the assessment, in terms of the content and construct validity (for an inference regarding a particular claim), an online administration can potentially provide a considerable advantage over other administration formats. Quite aside from the inferential course that the assessment traverses, the substance of the assessment argument can extend to both student-model and observable variables that are difficult to address with static and paper-and-pencil modalities of testing (Behrens et al., 2004; Williamson et al., 2004).

As we describe more and more complex assessment systems, however, the emphasis in this presentation will be placed on features of those assessments for which an online administration is recommended to enhance assessment argumentation. Because the use of innovative task types is a potential advantage of online administration in even this, the most basic of assessment systems, it is also a potential advantage of online assessments for *all* assessment systems discussed in this paper. Though the use of innovative task types will not be listed under each classification in the taxonomy, the reader should note that the potential benefits in employing an online administration that supports such task types applies to all cases.

Cell 2. *Fixed, examinee-controlled observations*. Traditional assessment with fixed, examiner-controlled claims and tasks affords little opportunity for the examinee to control the observations. Examples of this type include an examinee's freedom to work on different tasks in the same test section, choose the order to proceed through a section, revisit responses, and decide to omit responses to some tasks. These affordances do not play major roles in the examiner's reasoning, but they do introduce intertwined positive and negative effects: They provide the examinee with flexibility to achieve a higher level of performance if they are used well, but at the same time introduce construct-irrelevant variance among examinees to the extent they are not used well.

Cell 3. *Adaptive, examiner-controlled observations.* This class is similar to the traditional assessments in that the inferences are made about the same student-model variables for all examinees, and the examiner (or in many cases, a proxy for the examiner) controls the tasks. In contrast to traditional assessments, the observables are not constant across examinees. Examinees do not necessarily see the same tasks, and those that do might not see them in the same order. In other words, the frames of discernment the examiner works through with different examinees do not evolve with regard to student-model variables, but they do evolve, in some optimal manner, with regard to observable variables.

The most common example of this type of assessment is univariate IRT-based CATs (Hambleton & Swaminathan, 1985; Wainer & Mislevy, 1990). In these assessments, a task is administered and the response, typically in conjunction with an initial estimate of the student-model variable, is used to update the estimate of the student-model variable. The next task then is selected from among the available tasks, is administered, and leads to a revised estimate of the student-model variable. Algorithms for updating the student-model variable vary and may be based on maximum likelihood or Bayesian procedures (Wainer et al., 1990). Likewise, selection of the next task may be based on many features in addition to the current estimate of the student-model variable (e.g., task characteristics, frequency of task administration). Typically, the task to be presented next is (at least in part) a function of the current estimate of student-model variable and the psychometric properties of the tasks. For example, a task is selected that provides maximum information at the point of the current estimate of the student-model variable. Similarly, in a Bayesian framework, the task is selected on the basis of minimizing the expected variance of the posterior distribution for the examinee's student-model variable. In this way, the assessment is examiner controlled (via a proxy), and tasks are presented adaptively to facilitate better measurement in that the tasks any one examinee encounters are ideal or near-ideal. The result is greater measurement precision for a fixed amount of testing time and reduced bias in estimates of the student-model variable (Lord, 1983; Samejima, 1993).

Not all fixed-claim, adaptive-observation, examiner-controlled assessments employ IRT or require online administration. A century ago, Binet developed tests that called for the examiner to adapt the tasks long before the development of computers or IRT. To take another example, a five-question attitude survey may direct examinees that answered positively to Question 1 to respond to Questions 2 and 3, while directing examinees that answered negatively to Question 1 to respond to Questions 4 and 5. Such an assessment could be administered via paper and pencil as well as online. Important distinctions between these examples of examiner-controlled, adaptive assessments involve the number of observation adaptations (i.e., one for the attitude survey vs. many for item-level CAT), the number of tasks in the pool (i.e., five in the attitude survey vs. thousands for item-level CAT), and concern for test security (i.e., whether examinees will have access to tasks other than those that they are asked to complete). It is not feasible to present an examinee with a booklet of thousands of tasks and then direct them to respond to different tasks on the basis of their set of responses, particularly in high-stakes assessments where task security is a concern. We note these various possibilities and emphasize that the larger the task pool, the more adaptations required (both of which may be related to size of examinee pool), and the greater the concern for test security, the less and less feasible static assessments become.

Though the most common applications involve univariate IRT as a measurement model, the claim space may be multivariate (Segall, 1996). In a fixed, multiple-claims assessment with examiner-controlled adaptive observations, tasks are administered to provide observable variables that serve to update values of student-model variables that address the claims of interest. The assessment starts out focusing on a claim of interest and its associated student-model variable (for simplicity, assume there is a one-to-one relation between claims and student-model variables). Tasks are presented and observations are collected to statistically update the student-model variable; the tasks are presented adaptively, namely, on the basis of (at least) the current student-model-variable estimate and the psychometric parameters of the task. At some point, the assessment shifts focus to another claim and its associated student-model variable. As above, tasks are presented adaptively to update the estimate of the student-model variable, until at some point, the assessment shifts to another claim. This continues until the final claim is addressed, and the assessment concludes. As in the univariate-claim-space situation, the examiner controls the selection of subsequent tasks, which may vary over examinees.

In addition, the point at which the assessment shifts from one claim to another is also controlled by the examiner. Options for determining such a point include shifting the focus when: (a) the tasks appropriate for a particular claim are exhausted, (b) a predetermined level of statistical precision in the estimate of the student-model variable is reached, or (c) a certain time limit has been reached. The decision to shift may be influenced by the purpose of assessment and the particular claims involved. For high-stakes claims and decisions (e.g., will hiring this applicant make the firm vulnerable to an expensive lawsuit?), greater precision may be required for the relevant student-model variables before a shift in the assessment is warranted.

After the shift to a different claim, we are again faced with a decision regarding the initial task. In addition to the options discussed earlier, selection of the initial task for the second (or subsequent) claim might be informed by the examinee’s performance on earlier tasks. For example, if the examinee has performed well on the set of tasks pertaining to the first claim, and there is reason to believe the skills involved with the claims are positively related, the initial task for the second claim might be more difficult than if the examinee performed poorly on the first set of tasks.

A number of existing large-scale assessments have fallen into this category. For example, the Graduate Records Examinations General Test (Mills, 1999) consisted of verbal, quantitative, and analytical sections. Though the total claim space was multidimensional, unidimensional IRT was employed in each section. Tasks for each section informed upon a student-model variable localized to the particular section. For each section, the examiner, who also controlled the shifting and the stopping of the assessment, adaptively selected the tasks.

Adaptive IRT allows for higher ability examinees to be given tasks suitable for them; they are not presented too many easy tasks that may lead to boredom or carelessness. Likewise, lower ability examinees are given tasks suitable for them; they are not presented with too many difficult tasks that may lead to frustration or an ability estimate that is influenced by lower ability examinees' tendencies to guess on inappropriately difficult tasks. The details of these and other mechanisms for task selection are beyond the scope and intent of this paper. For current purposes, it is sufficient to note that the necessary calculations involved in estimating student-model variables and shifting the assessment focus, even when approximations to such calculations are employed (e.g., the use of information tables), are computationally intensive enough that they require a computer.

Cell 4. *Adaptive, examinee-controlled observations*. In contrast to the examiner-controlled adaptive assessments just described, this family of assessments permits the examinee to select the tasks—or given a fixed initial task, permits the examinee to select subsequent tasks—on the fly. The frame of discernment does not evolve with regard to the student-model variable(s), which is (are) fixed and controlled by the examiner, but it does evolve with respect to observable variables, in a manner controlled by the examinee. This shared responsibility for the evolution of the frame of discernment immediately raises the question of the principles on which tasks are selected. As mentioned above, rules for examiner-controlled adaptive observations involve comparisons of the tasks. Implicitly, knowledge of the task properties is required; task selection algorithms typically involve maximizing information or minimizing expected posterior variance regarding the student-model variable(s). Furthermore, these algorithms are often subject to constraints regarding task content, structure, exposure, and so forth. Without question, it is unreasonable to demand examinees to make such decisions on these criteria on the fly as the decisions involve overly burdening computation. What's more, setting aside the properties of the tasks, selecting in this manner requires being aware of all the tasks. Though examinees are often familiar with types of tasks (especially in large-scale, high-stakes assessments), it is not the case that they have seen all the tasks from which to select. Clearly, if examinee-controlled, adaptive-observation assessments are to exist, they are to have a considerably different essence than that of the examiner-controlled, adaptive-observation assessments. In what follows, we describe two flavors of examinee-controlled, adaptive-observation assessments for a fixed-claim space.

Consider an assessment where tasks are developed to provide evidence for a single claim. Suppose, as occurs in assessments in a number of disciplines at the undergraduate and graduate levels, the examinees are presented with all the tasks and informed as to the order of difficulty of the tasks and how their work will be evaluated. A natural scoring rule would have a correct observable worth more than an incorrect observable, and harder observables would be worth more. For example, Wright (1977) described a self-adaptive test in which a student chooses items one page at a time from a relatively short test booklet, scoring is based on the Rasch model, and correct responses to harder items induce likelihoods that are peaked at higher levels of the latent ability variable. The examinee then selects a finite number of tasks to complete and submit. Examinees will then not necessarily have values on the same observable variables; each examinee individually determines which variables will have values. Such an assessment model is easily generalizable to multiple claims.

This example illustrates the case where an assessment is adaptive (in terms of the observations), and yet there is no feedback provided to the examinee for the process of selecting the next task. The assessment is adaptive in the sense that the frame of discernment varies across examinees. Each examinee's frame of discernment, the collection of student-model and observable variables, is not fixed a priori, but rather is determined as the assessment unfolds.

There are two concerns with this type of examinee-controlled, adaptive testing—one practical and the other statistical. Practically, such assessments would have to consist of a task pool small enough for the examinees to review and select from among all the tasks, and the assessments would not be appropriate if task security was a concern. Statistically, care would need to be taken to avoid the bias incurred by not-answered questions that Rubin (1976) called nonignorably missing. (A simple example is filming yourself attempting 100 basketball free throws, making 20, and editing the film to show the completed baskets and only five misses.) This can be accomplished by allowing choice among items that differ as to ancillary knowledge, but all demand the same targeted knowledge. For example, an examiner can ask for a Freudian analysis of a character in a Shakespearean play and let students choose a play that was familiar to them. This focuses evaluation on the Freudian analysis, while assuring familiarity with the character.

Another type of fixed-claim, examinee-controlled, adaptive-observation assessment is self-adaptive testing, a variant of more familiar (examiner-controlled) CAT. To date, all self-adaptive tests (SATs) have employed IRT to achieve adaptation. In SATs (Rocklin & O'Donnell, 1987; Wise et al., 1992), tasks were grouped into a finite number (typically six or eight) of bins based on difficulty, namely the *b* parameter in IRT. Upon completion of each task, examinees choose how difficult the next task will be by choosing among the bin from which the next item would be selected. Once the examinee selected the difficulty level, a task from that bin may be selected randomly or on the basis of maximizing information. The latter case represents a hybrid of examiner- and examinee-controlled assessments. Input is necessary from both agents to select the next task.

Similarly, other considerations lead to hybrid assessments. In cases where the examinee repeatedly selects tasks from one difficulty bin, the examinee may exhaust the tasks in that bin before the assessment is complete (Wise et al., 1992). Or, if the selected bin is far from an examinee's ability level, ability estimates will be biased (Lord, 1983; Pitkin & Vispoel, 2001; Samejima, 1993). To control for these possibilities, the task selection algorithm may be constrained so that examinees are forced to select tasks from different bins, particularly if they are repeatedly correct (or incorrect) in their responses to tasks from a particular bin (Vispoel, 1998). Again, such an alteration results in a hybrid of examiner- and examinee-controlled assessment.

Several studies have shown that SATs can lead to reduced test anxiety and higher ability estimates, as compared to examiner-controlled CATs (e.g., Rocklin & O'Donnell, 1987; Wise et al., 1992), though some studies have found these effects to be negligible or nonexistent (for a review, see Pitkin & Vispoel, 2001). Several theories for how SATs might counter the effects of test anxiety on performance exist. See the discussion in Pitkin and Vispoel (2001) and the references therein for a full review.

What is of greater concern in this work is an understanding of the convergent and divergent aspects of examinee-controlled SATs and the more traditional examiner-controlled adaptive-observation tests and the implications for test use. As Vispoel (1998) noted, the potential advantage of reducing construct-irrelevant variance (e.g., anxiety) via SATs does not come without a price. In particular, there is a loss in precision, as standard errors of ability estimates are higher for SATs (Pitkin & Vispoel, 2001; Vispoel, 1998), and a loss in efficiency, as SATs require more time (Pitkin & Vispoel, 2001). This result is to be expected when we recognize that examiner-controlled CATs are built to maximize precision. To the extent that the tasks selected deviate from those that would result in maximum precision (as will almost surely be the case in SATs), there will be a loss in the precision, or, in the case where the stopping criterion is based on precision of the estimate of the student-model variable, an increase in testing time.

In terms of use, we follow Pitkin and Vispoel (2001) in noting that possible bias, loss of precision, sensitivity to test-wiseness, and increased costs in item-pool development and management are some of the difficulties involving the use of SATs in high-stakes assessments. Further, we follow Pitkin and Vispoel (2001) in lamenting the fact that the effects of reducing test anxiety might be most pronounced and desirable in high-stakes assessments. Nevertheless, SATs may be appropriately used for low-stakes diagnostic purposes. In particular, SATs with feedback (Vispoel, 1998) may offer ideal properties for diagnostic assessments. Feedback given to examinees may be as simple as whether they completed the task correctly and may aid the examinee in selecting a task bin that is more appropriate (i.e., closer to their ability level), which would result in observed increase in precision in SATs with feedback vs. those without (Vispoel, 1998). An SAT with feedback is a step in the oft-desired but rarely achieved direction of an integration of assessment and instruction via a computer-based assessment system (Bunderson et al., 1988).

Reporting whether the task was completed correctly only scratches the surface of the level of feedback that may be given. That is, if the tasks are constructed appropriately, features of the work product (above and beyond "right" or "wrong") may serve as evidence regarding the examinee's cognitive abilities. This may be the case even if the task is as simple as the selection of a particular option in a multiple-choice question. For example, when solving problems in physics, students may employ principles derived from Aristotle, Newton, or Einstein (among others). If distractors are constructed to be consistent with incorrect frames of thinking, then the selection of those distractors by an examinee might be able to pinpoint the extent to which the examinee understands (or fails to understand) the relevant principles of physics. Such information would be relevant to examinees in a diagnostic setting or to examiners in both diagnostic and summative settings; an online administration permits immediate feedback to examinees and examiners.

Extensions to multiple-claims assessments are straightforward. The assessment commences with a task regarding one claim and the examinee selects subsequent tasks after completing each one. At some point, the assessment shifts to tasks that provide evidence for another claim, and the same process occurs. After completing an initial task (initial to this new claim), the examinee chooses the difficulty bin for the next task. Theoretically, there is no limit on the number of claims that can be addressed in this way.

Several decisions, some more implicit than others, are necessary in administering such an assessment. A number of options exist for selection of the first task. Because examinees will choose between harder and easier tasks, a sensible choice would be somewhere in the middle of the difficulty distribution. Alternatively, one could start with a comparably easier task, with the expectation that most examinees will then opt for a more difficult task.

In the case of multiple-fixed claims, specifying the change point and the initial task for a new claim may be accomplished in a number of ways, as discussed in the preceding section. With all the computation involved in selecting an initial task, accepting examinee input in terms of the bin to use, and selecting a task from the bin (either randomly or to maximize information), even the simplest SAT can only be administered online. With the increased complexity in hybrid algorithms for task selection and, in the case of multiple-claim assessments, shifting the focus to "another claim"—particularly when the shift is based on an achieved level of precision—the need for an online administration becomes even more evident.

Fixed, Examinee-Controlled Claims

Akin to the situation in Cell 2, it makes little sense to say the claims are fixed, and therefore not subject to evolve over the course of the assessment, and yet controlled by the examinee. This reasoning applies to Cells 5, 6, 7, and 8 in the taxonomy. In passing, it is noted that Cell 6 states that both claims and observations are fixed yet controlled by examinees and is, therefore, doubly nonsensical.

Adaptive, Examiner-Controlled Claims

In adaptive claim assessments, the claim space is necessarily multidimensional. All the classes of assessments discussed in this and subsequent sections are adaptive and hence multidimensional. They are adaptive in the sense that the inferences drawn may vary across examinees. The assessments are examiner controlled in the sense that the choice of the claim to consider and the choice of when to move to another point in the claim space are controlled by the examiner, rather than the examinee. The distinctions among these assessments, discussed in the following subsections, have to do with the status of the observations.

Cell 9. *Fixed, examiner-controlled observations.* This class of assessments is defined by examinees responding to the same tasks, the selection and presentation of which are in control of the examiner, while the inferences drawn vary across examinees. That is, examinees all encounter the same tasks, but the inferences drawn may be at different points in the claim space. An example of this includes analysis of a Rorschach test in which examinees are all presented with the same stimuli, but the responses lead the clinician to create an individualized interpretation that can involve different claims for different examinees.

Another example may be drawn from the Minnesota Multiphasic Personality Inventory–2, or MMPI–2 (Butcher et al., 1989). An examinee taking the full MMPI–2 sees hundreds of tasks that are fixed and examiner controlled. The examiner may then form different scales from these, adapting what is formed in light of the examinee. Though the observations are fixed, the frame of discernment alters as the claim of interest changes.

Two features of this type of assessment are noteworthy. First, as discussed above, that a claim space can be adaptive indicates that it is multidimensional. Second, given that they are multidimensional, fixed-observation assessments are in many cases inefficient. If the claim space is multidimensional and fixed, an appropriate number of tasks can be constructed and selected a priori for each claim (in which case it will be a fixed-claim, fixed-observation assessment described in Cell 1 or the tasks can be selected on the fly (i.e., a fixed-claim, adaptive-observation assessment described in Cell 3). However, if the claim space is multidimensional and adaptive, a part of the goal is to allow the assessment to adjust the focus—the inferential target—during the assessment. Since observables that are optimal for certain claims are most likely not optimal for other claims, moving around the claim space adaptively calls for the selection of the observables to be adaptive as well. We take up adaptive-claim, adaptive-observation assessments in subsequent sections.

Cell 10. *Fixed, examinee-controlled observations.* As in the discussion of Cell 2, it offers little to an understanding of the analysis of argumentation to dwell on those marginal situations in which the observations are fixed, yet controlled by examinees.

Cell 11. *Adaptive, examiner-controlled observations.* In an assessment where summative inferences may be sought for multiple claims, an adaptive-claim, adaptive-observation assessment with examiner control of both claims and observations is ideal. To introduce this type of assessment, we begin by generalizing the more familiar fixed-claim, examiner-controlled, adaptive-observation assessments (see Cell 3).

In Cell 3, fixed-claim, adaptive-observation assessments were discussed, and common procedures for adapting the observations were mentioned. The main purpose of adapting is to provide an assessment that is optimal for each examinee. Implicit in the discussion was the constraint that the inferences to be made were, for all examinees, with regard to the same claim(s). In adaptive-claim assessments, this constraint is released; the inferences made from an adaptive-claim assessment may vary across examinees, not only in their values (i.e., this examinee is proficient in math, this examinee is not proficient in math), but in the variables as well.

Results from the assessment might lead to inferences for an examinee regarding proficiency in one area of the domain (with an associated claim or set of claims), while inferences for another examinee would concern proficiency in a *different* area of the domain (with its own separate claim or claims). As an examinee proceeds through the assessment, evidence is gathered. As evidence is gathered, certain hypotheses are supported while others are not, which leads to questions about other hypotheses; these questions may differ between examinees. In fixed-claim, adaptive-observation assessments, the evidence differs between examinees, but the inferential question asked is the same. In adaptive-claim assessments, the inferential questions differ as well.

For example, consider an assessment in which tasks are constructed such that examinees may employ one of possibly several cognitive strategies in approaching or solving the tasks. The assessment could then adapt the claims on the basis of examinee performance. If performance on tasks early in the assessment indicates the examinee is employing a particular strategy, the assessment claim can be defined or refined to focus on that strategy, and tasks may be adapted accordingly, so to provide maximum information regarding that claim for that examinee. Another examinee, employing a different cognitive strategy, will have the assessment routed to focus on a claim regarding that strategy and will encounter appropriate tasks to obtain evidence for that claim. For both examinees, as information regarding a particular claim is incorporated, new questions regarding other claims may result. The assessment then shifts to address those claims, adaptively administering tasks to provide observable evidence regarding student-model variables for those claims. This process continues until the end of the assessment. Though the assessment may be broadly labeled with a general term, the results of the assessment will yield different inferential targets.

For example, a developing line of research has investigated the cognitive strategies employed by students in tackling problems of mixed number subtraction (de la Torre & Douglas, 2004; Mislevy, 1996; C. Tatsuoka, 2002; K. Tatsuoka, 1990). Under one strategy, a set of attributes is necessary to successfully complete the tasks, while under another strategy, a different (though possibly overlapping) set of attributes is necessary. One could devise an assessment that seeks to identify which strategy an examinee is employing in addressing the problems at hand and then select tasks that are most informative for that particular strategy. For examinees choosing a particular strategy, the assessment provides information relevant to claims associated with the attributes necessary for that strategy; it cannot speak to claims associated with attributes that are not part of that strategy. Though the assessment may be broadly labeled “mixed number subtraction,” the actual inferential targets vary over examinees on the basis of their cognitive strategies.

As argued earlier, if the observations are to be adapted between examinees, an online administration is all but required. All the computational complexity is increased when both the claims and the observations are free to vary between examinees. Facilitation of individualized inferences using optimally selected tasks can only be accomplished via an online administration.

Cell 12. *Adaptive, examinee-controlled observations.* These assessments might be thought of as slight changes to either the assessments described in Cell 11 or Cell 4. Similar to those in Cell 11, these assessments involve multiple claims that are controlled by the examiner. In Cell 11, the examiner adapts the observations. Here, the observations are adaptive but controlled by the examinee. Likewise, in Cell 4, the examinee controlled the observations related to a fixed (set of) claim(s) set out by the examiner. Here, the examinee controls the observations; the claims, though still controlled by the examiner, vary over examinees.

Recognizing that Cell 11 builds off the CATs described in Cell 3 by permitting there to be multiple claims, and that Cell 4 builds off the CATs described in Cell 3 by granting control of the observations to examinees, the current category can be seen as the combination of those changes. The focus of the assessment, though controlled by the examiner, varies over examinees; the observations also vary, as determined by the examinees. In a sense, these assessments are SATs with multiple claims that are controlled by examiners. The features, benefits, and drawbacks of examinee-controlled observations (see Cell 4) and examiner-controlled adaptive claims (see Cell 11) are combined.

Again, suppose tasks have been constructed such that examinees may employ one of possibly several cognitive strategies in approaching or solving the tasks. The assessment could then control the claims on the basis of examinee performance, all the while permitting examinees to have input into what tasks (within the family of tasks for that claim) are selected. If performance on tasks early in the assessment indicates the examinee is employing a particular strategy, the assessment claim can be defined or refined by the examiner to focus on that strategy, while the difficulty of the tasks would be controlled by the examinee, say by binning items and prompting the examinees for which bin to select from, as in conventional SATs.

Recent advances in intelligent tutoring systems include the development of innovative assessment models to support intelligent tutoring customized to the examinee’s knowledge and problem solution strategy. Andes, an intelligent tutoring system for physics (Gertner & VanLehn, 2000), dynamically builds student models as the student proceeds through the tasks. Once a student selects a task, Andes loads the solution graph, a network representation of the relevant knowledge, strategies, and goals involved in successfully solving the problem. The solution graph is automatically converted into a student model in the form of a Bayesian network (Conati et al., 1997; for more on Bayesian networks, see Almond & Mislevy, 1999; Jensen, 2001; Martin & VanLehn, 1995; Mislevy, 1994; Pearl, 1988). For each task in Andes, there is a Bayesian network containing nodes for all the relevant facts, rules, strategies, and goals. As the student solves the task, nodes may be fixed to certain values, other nodes may be added dynamically, and others may be updated in accordance with what the student does via propagation of evidence through the network.

Once the student selects a new task, the nodes relevant to the old task are discarded and the nodes relevant to the new task are added. Nodes relevant to both tasks are retained. In this way, the information from previous tasks is brought, so subsequent tasks—the state of the nodes after the previous task—become the prior distribution and initialize the model for the new task. Over the course of the assessment, as evidence regarding student knowledge of facts, familiarity with rules, and use of strategies enters the model, the assessment automatically moves around the claim space. In addition to the values of the student-model variables being updated, the contents of the student model—the variables themselves—change as beliefs about the student’s knowledge, abilities, strategies, and goals change.

In Shafer's terms (1976), the frame of discernment adapts on the fly for each examinee as they proceed throughout the system. From task to task, the student model changes, and information regarding the examinee addresses some hypotheses and brings to light others that remain to be addressed.

What is key for the current purpose is recognizing that the additional complexity of adaptive claims, moving throughout the claims space and adjusting the target of inference, essentially requires an online administration. In addition to the computational requirements for storing and presenting the various tasks, the adaptation of the claims also depends on computational power.

Adaptive, Examinee-Controlled Claims

This class of assessments differs from all those previously discussed in that the examinee is in control of the target of inference. As an adaptive claim assessment, the claim space is multidimensional and the inferences may vary across examinees. Here, the decisions regarding which claims to address and when to shift from one claim to another is in control of the examinee.

Cell 13. *Fixed, examiner-controlled observations.* Assessments of this sort may be described as slight changes to those in Cell 9. Recall the example of the MMPI–2, in which an examinee encounters hundreds of tasks that are fixed and examiner controlled. In Cell 9, the exploration of the scales that can be formed was controlled by the examiner. Here, the examinee chooses the scales to explore. As in Cell 9, having a fixed set of observations may be inefficient for adaptive-claims assessments.

Cell 14. *Fixed, examinee-controlled observations.* As in sections discussing Cells 2, 6, and 10, little is gained toward the end of explicating the structures of assessment arguments by consideration of those situations in which the observations are fixed yet controlled by the examinee.

Cell 15. *Adaptive, examiner-controlled observations*. These assessments might be thought of as slight changes to the assessments described in Cell 11. Similar to Cell 11, the observations vary between students and are obtained based on examiner-controlled task presentation. In addition, the claims may vary between examinees. In contrast to Cell 11, control of the claims is in the hands of the examinee. Thus, the focus of the assessment is controlled by the examinee. In short, the examinee chooses the target of interest (i.e., the claim) and then the examiner controls what tasks are presented. The assessment is ideally suited for diagnostic assessments in which the examinee determines the area of focus, say, regarding certain areas in which the examinee would like some feedback concerning their achievement level. Once the focus is determined, the examiner presents tasks to obtain maximal information employing the methods already described.

Again, the complexity involved with having libraries of tasks relevant to possibly many claims only adds to the computational requirements of adapting the tasks on the basis of previous performance. As with simpler assessments that involve adapting in simpler ways, any large-scale application is feasible only with an online administration. The assessments described here are well-suited for diagnostic purposes under the guidance of each examinee. As such, possible situations for employing these systems are longitudinal diagnostic assessments. In the course of an instruction period, students could engage in the assessment, selecting the focus of the assessment while the examiner selects the most appropriate tasks. At a later time, the examinee could engage with the assessment system again; selection of the same claim(s) would lead to current estimates of the examinees' proficiencies with regard to that claim. This provides a natural way for the student to track their own progress over time.

Although we are not aware of any educational assessments in this cell, there is an analogue in Internet sites that helps people explore what cars, careers, books, or movies they might like (e.g., the SIGI^3 career planner; see http://sigi3.org/). Standard questions about what the user likes to do, what is important to the user, how the user makes proffered choices, and so forth help the user figure out classes or properties of cars, careers, books, or movies to investigate more deeply. With examiner-adaptive observations, answers to earlier questions can influence what questions will be asked next. One site for helping elementary school children find books they might like is Book Adventure (https://www.bookadventure.com). Of course, librarians also do this in person with students. The problem is that even though all the information is available in the library, it overwhelms young students. Only the students "know" what the ultimate claims of interest will turn out to be. A program's frame of discernment uses examiner-created observables and student-model variables, and as an interview proceeds, the frame of discernment is increasingly under the control of the student.

Cell 16. *Adaptive, examinee-controlled observations.* The final category consists of assessments that allow examinees to control both the claims and the tasks to yield observations for those claims. The examinee selects the claims to focus on and then has input into the observed data, for example in the manner of SATs described above.

Information-filtering and user-modeling systems involve these types of assessments of this class (e.g., Rich, 1979; this source is a bit outdated in terms of current cognitive theory, but the beginning is excellent in terms of laying out the situation as an inferential problem that is aligned with the taxonomy proposed here). For example, a central problem in information systems involves the retrieval systems in libraries that organize materials and search terms that try to help patrons find the information they might want, without knowing what it is that any new patron might want.

Consider a simple case where a user's query results in a list of documents, possibly structured by some criterion such as perceived relevance. The user then selects some of the documents from the list for further consideration. A great deal of observable information can be collected from such a process. Which documents were viewed? In what order? How much time did the user spend reading each? These only scratch the surface of what data could possibly be collected. In these systems, the user is in control of the claim space, via the query, and the observables, via the actions taken with respect to the produced list of documents.

A more refined example comes from NetPASS, a computer-based interactive assessment in the domain of computer networking containing tasks targeted toward the related but distinct aspects of network design, implementation, and troubleshooting (Behrens et al., 2004; Williamson et al., 2004). Upon selecting one of these areas of the domain, examinees select the desired level of difficulty (easy, medium, hard) of a task. Thus, the examinees control both the claims and observations. For each task, there is a Bayesian network fragment containing all relevant student-model and observable variables (Levy & Mislevy, 2004). As tasks are completed, values for the observables are entered and information is propagated throughout the network, updating the student-model variables. When a new task is called, the variables associated with the previous task that do not pertain to the new task are dropped, and previously unused variables relevant for the new task are included. In this way, variables are docked into or dropped out of the network, as needed (Mislevy et al., 1998). Here, as is the case with Andes, the use of a Bayesian network for information propagation supports the adaptation and the flexibility it provides.

In Cell 15, it was argued that an assessment in which the examinee controls the focus was more suited for diagnostic than summative assessment. Likewise, in Cell 4 it was argued that assessments in which the examinee controls the observations are likely to be inefficient for estimation of parameters pertaining to the claims and thus may be inefficient as summative assessments. Assessments in this final class combine the examinee-controlled features of Cells 4 and 15 and are ideally suited to diagnostic assessment. As with the other classes of assessments that involve adaptation of observations, the need for an online administration is clear. And, as in the classes that involve adaptation of claims as well as observations, the need for an online administration is increased.

Discussion

The focus of this work is to detail different ways an assessment system can operate in terms of the targets of inference and the tasks presented to examinees. The taxonomy described here classifies assessments in terms of the claim status, observations status, and the controlling parties. Well-known univariate IRT has been employed to facilitate both examiner-controlled and examinee-controlled, fixed-claim assessments. The advantages of an online administration, namely, high-speed computations regarding evidence accumulation and task selection, make adaptive-observation assessments feasible. More complex assessments involving adaptive claims have yet to achieve the prominence of adaptive-observation assessments.

We propose two reasons for this. First, the majority of traditional paper-and-pencil assessments were fixed-observation assessments. Limitations of fixed-observation assessments (e.g., inefficiency in terms of appropriateness of tasks) were known before the advent of online administration. Thus, the capabilities of an online administration were first used to combat these limitations via adapting the observations, rather than extending to multiple, adaptive claims. Second, in order for the examiner-controlled, adaptive-claims assessments described here to actually be effective, considerable work must be done up front. In the case of an assessment system that adapts to the examinee’s chosen strategy for solving subtraction problems, cognitive studies on the reasoning patterns employed by students must be done, and the tasks must be constructed and calibrated such that they are consistent with this cognitive work. This work will most likely need to be done domain by domain. Only recently has the cognitive groundwork necessary for such complex assessments been laid in certain domains (for an example in the domain of computer networking, see Williamson et al., 2004). In efforts to extend assessment in these directions, research and experience in the fields of user modeling in such domains as consumer preferences, adaptive software engineering, and information sciences should prove useful.

To summarize, adaptation enhances the validity argument for the assessment. This holds both for adapting the observations (e.g., increased measurement precision, decrease in bias, decrease in test anxiety) and adapting the claims (e.g., identification of cognitive strategies, individualized diagnostic feedback for both examiners and examinees). Assessment systems with adaptation all but require an online administration, especially for large-scale assessment. What is more, in providing increased security, decreased scoring errors, faster score reporting, and the opportunity for innovative task types, an online administration can be advantageous even in situations without adaptation.

No declaration is made about the taxonomy presented here being exhaustive. Already we have mentioned settings in which the locus of control for either the claims and/or the observations would be a hybrid of examiner- and examinee-controlled assessments. Further refinements in the future are eagerly anticipated. Nevertheless, framing assessments in terms of the observation status, claim status, and the locus of control for these aspects proves useful in (a) designing/aligning an assessment with the purpose at hand, (b) understanding what options are available in terms of assessment design and operationalization, (c) documenting strengths and weaknesses of assessments, and (d) making explicit the features of the assessment argument. Though not described here, the taxonomy also proves useful for designing or selecting an appropriate statistical measurement model. Future work in this area will include aligning various existing statistical models with the taxonomy and suggesting the possible advantages (and disadvantages) of both more complex statistical models and adaptive reconfigurations of simple models.

References

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 223–237. CrossRef

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5). WebLink

Baker, E., Dickieson, J., Wulfeck, W., & O'Neil, H. F. (Eds.). (2017). *Assessment of problem solving using simulations.* Routledge. CrossRef

Behrens, J. T, Collison T. A., & DeMark, S. (2006). The seven C's of comprehensive online assessment: Lessons learned from 36 million classroom assessments in the Cisco Networking Academy program. In S. L. Howell & M. Hricko (Eds.), *Online assessment and measurement: Case studies from higher education, K-12 and corporate* (pp. 229–245). IGI Global. CrossRef

Behrens, J. T, Mislevy, R. J., Bauer, M., Williamson, D. W., & Levy, R. (2004). Introduction to evidence-centered design and lessons learned from its application in a global e-learning program. *International Journal of Testing*, *4*(4), 295–301. CrossRef

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The four generations of computerized testing. In R. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 367–407). Macmillan.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory–2 (MMPI–2): Manual for administration and scoring.* University of Minnesota Press.

Conati, C., Gertner, A. S., VanLehn, K., & Druzdzel, M. J. (1997). Online student modeling for coached problem solving using Bayesian networks. In A. Jameson, C. Paris, C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97* (pp. 231–242). Springer. CrossRef

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. CrossRef

Gertner, A., & VanLehn, K. (2000). Andes: A coached problem-solving environment for physics. In C. Frasson, G. Gauthier, & K. VanLehn (Eds.), *Intelligent tutoring systems–5th international conference, ITS 2000 proceedings* (pp. 131–142). Springer. CrossRef

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Kluwer-Nijhoff. CrossRef

Ifenthaler, D., & Kim, Y. J. (Eds.). (2019). *Game-based assessment revisited.* Springer. CrossRef

Jensen, F. V. (2001). *Bayesian networks and decision graphs*. Springer. CrossRef

Ke, F., Shute, V. J., Clark, K. M., & Erlebacher, G. (2019). *Interdisciplinary design of game-based learning platforms: A phenomenological examination of the integrative design of game, learning, and assessment.* Springer. CrossRef

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing*, *4*(4), 333–369. CrossRef

Lord, F. M. (1983). Unbiased estimators of ability parameters, their variance, and their parallel-forms reliability. *Psychometrika, 48*(2), 233–245. CrossRef

Martin, J. D., & VanLehn, K. (1995). A Bayesian approach to cognitive assessment. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 141–165). Lawrence Erlbaum.

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Records Examinations General Test. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 117–135). Lawrence Erlbaum.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*(4), 439–483. CrossRef

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379–416. CrossRef

Mislevy, R. J., Almond, R. G., & Steinberg, L.S. (1998). *A note on the knowledge-based model construction in educational assessment* (CSE Technical Report No. 480). University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). WebLink

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62. CrossRef

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Kaufmann. CrossRef

Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement, 38*(3), 235–247. CrossRef

Rich, E. (1979). User modeling via stereotypes. *Cognitive Science, 3*(4), 329–354. CrossRef

Rocklin, T. R., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79*(3), 315–319. CrossRef

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. CrossRef

Samejima, F. (1993). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika, 58*(2), 195–209. CrossRef

Schum, D. A. (1994). *The evidential foundations of probablistic reasoning*. Wiley-Interscience.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*(2), 331–354. CrossRef

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press. CrossRef

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C (Applied Statistics), 51*(3), 337–350. CrossRef

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Lawrence Erlbaum.

von Davier, A. A., Mislevy, R. J., & Hao, J. (Eds.). (2021). *Computational psychometrics: New methods for a new generation of educational assessment*. Springer. CrossRef

von Davier, A. A., Zhu, M., & Kyllonen, P. C. (Eds.). (2017). *Innovative assessment of collaboration*. Springer. CrossRef

Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and test anxiety. *Journal of Educational Measurement*, *35*(2), 155–167. CrossRef

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (Eds.). (1990). *Computerized adaptive testing: A primer*. Lawrence Erlbaum.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration and proficiency estimation. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 65–102). Lawrence Erlbaum.

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*(4), 303–332. CrossRef

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1992). A comparison of self-adapted and computerized adaptive tests. *Journal of Educational Measurement*, *29*(4), 329–339. CrossRef

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116. CrossRef

Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). (2020). *Handbook of automated scoring: Theory into practice*. CRC Press. CrossRef

Author Address

Robert J. Mislevy. *Email*: rmislevy@umd.edu