The Gordon Commission
on the Future of Assessment in Education

Assessment as Evidential Reasoning

Joanna S. Gorin
Arizona State University

Introduction

As defined by the 1999 *Standards for Educational and Psychological Testing* (APA/AERA/NCME, 1999), an assessment is "any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects, or programs". It further defines a variety of types of assessment (e.g., cognitive assessment, psychological assessment, attention assessment, standards-based assessment, performance assessments, and vocational assessment). For example, cognitive assessment is defined as "the process of systematically gathering test scores and related data in order to make judgments about an individual's ability to perform various mental activities involved in the processing, acquisition, retention, conceptualization, and organization of sensory, perceptual, verbal, spatial, and psychomotor information." Psychological assessment is defined as "a comprehensive examination of psychological functioning that involves collecting, evaluating, and integrating test results and collateral information, and reporting information about an individual. Various methods may be used to acquire information during a psychological assessment: administering, scoring and interpreting tests and inventories; behavioral observation in multiple contexts (e.g., classrooms, home); client and third-party interviews; analysis of prior educational, occupational, medical, and psychological records." What is notable about each of these definitions is that each remarks on the use of multiple sources of data, information, or evidence across varied contexts to be used in drawing conclusions. Despite these accepted definitions, current educational assessments are most commonly based on administration of a single objectively-scored context-free standardized test. In fact, this practice is so widespread that the terms "educational assessment" and "educational test" are often used interchangeably. Such practice assumes at least one fact to be true – it is possible to capture all relevant information to answer our assessment questions at a single point in time and in a single context. This assumption directly contradicts the premise of APA's standards for assessment, that evidence must be gathered from multiple sources in multiple contexts at multiple times.

The future of educational assessment requires an expanded framework that encompasses all relevant opportunities to gather evidence in support of our assessment purpose. In particular, evidence from contextually assessment-tasks that incorporate the complex of variables and factors likely to impact real-world behavior will be most useful for purposes of generalizability of score interpretation. By broadening our definition of assessment and evidence to better align

with the multi-dimensional and contextually rich definitions of *The Standards*, educational assessments will become more relevant to the educational and instructional processes than at present. This requires a fundamental change to the way in which we view assessment purpose and assessment design. One possible conceptualization is that of assessment as evidential reasoning.

Assessment as Evidentiary Arguments and Evidential Reasoning

Mislevy (2012) introduced four metaphors of assessment, one of which was the metaphor of assessment as an evidentiary argument about students' learning and abilities given their behavior in particular circumstances. As in all evidentiary reasoning, the quality and persuasiveness of the argument is primarily a function of the evidence used to support the argument. The quality of an assessment argument is therefore a function of the evidence gathered in the process of assessment and its quality. This paper examines limitations of the dominant practice in educational assessment of the 20th century, which typically reduces the assessment argument to a single piece of evidence – standardized test scores. I explore here a more expansive view of educational assessment as an evidentiary argument that draws from multiple evidential sources in order to make the most valid and reliable claims about student learning. I argue that assessments based on multiple evidence sources from contextually rich situated learning environments, including unconventional data regarding student engagement, motivation, opportunity-to-learn, and socio-cultural experiences, offer an expanded definition of assessment that will improve the ability to make valid decisions about student learning and instruction. Data not previously considered as part of assessment arguments will permit claims about skills, attitudes, behaviors, and temperaments not previously considered relevant for educational assessment. Further, they may be considered as part of our traditional assessment arguments by altering our interpretation of cognitive abilities data, adding qualifications or alternative hypotheses to explain student behavior.

Evidential Reasoning

Under conditions of uncertainty, decisions should be made based on evidence that makes certain alternatives more or less likely than one another. The evidential reasoning approach (ER) is a quantitative approach newly used in decision theory to make informed decisions based on

available evidence (Xu, Yang, & Wang, 2006; Yang & Xu, 2002). The ER process relies heavily on evidentiary arguments that can be formalized in terms of decision alternatives and evidence-based criteria. ER's origins are rooted in Stephen Toulmin's (1958) framework and terminology for analyzing arguments (See Figure 1). Six critical terms are used to structure the argument:

1. *Claim*: the position or claim being argued for; the conclusion of the argument.
2. *Grounds:* reasons or supporting evidence that bolster the claim.
3. *Warrant:* the principle, provision or chain of reasoning that connects the grounds/reason to the claim.
4. *Backing*: support, justification, reasons to back up the warrant.
5. *Rebuttal/Reservation:* exceptions to the claim; description and rebuttal of counter-examples and counter-arguments.
6. *Qualification*: specification of limits to claim, warrant and backing. The degree of conditionality asserted.

The grounds are data collected as evidence that either support or refute the desired claim. The warrant justifies the use of the data by virtue of backing, which illustrates the data's meaning and usefulness. Exceptions and limitations of the inference to specific situations can be described through rebuttals and qualifications. The strength of the inference lies primarily in a) the amount and quality of the data, b) the strength of the backing to support the warrant. Together these constitute the argument's evidence. The entirety of the argument is evaluated in terms of the likelihood of the claim based on the evidence presented.

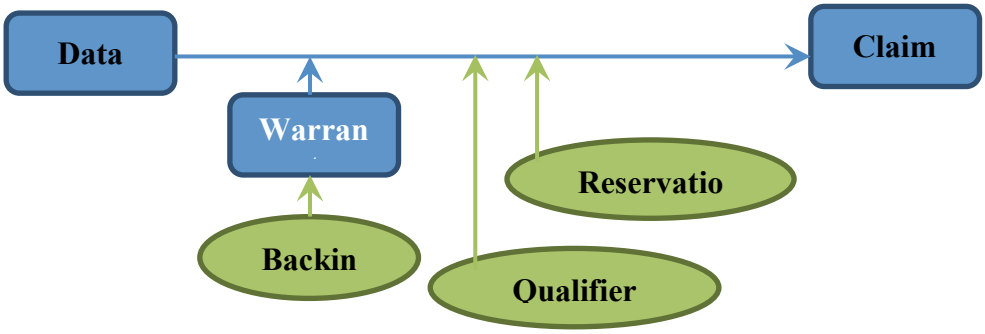

Figure 1. Toulmin's model of inference and argument structure.

The Assessment Argument

How does evidential reasoning pertain to educational assessment? To paraphrase the definitions given by *The Standards*, the process of educational assessment is to synthesize data from multiple disparate sources of evidential information to make claims about the knowledge, skills, attitudes, and beliefs of individual students as individuals or aggregated groups. The score interpretation or assessment purpose is the claim, the data is the student behavior, and the warrant and backing are additional information about the item and the behavior of the student (e.g., existing research and theory about item performance and student cognition). It seems clear that by definition alone, an assessment is a form of an evidential argument. Having accepted this claim, what can further be gained by formally framing educational assessment as an argument?

Principles of evidential reasoning have often been discussed in the context of educational and psychological measurement with respect to construct validity and validity arguments (Cronbach, 1989; Kane, 1992; Messick, 1989). More recently, Mislevy (1994) featured the importance of evidence throughout the entire assessment design and development process in an assessment design framework called *Evidence Centered Design (ECD)*. An ECD approach to educational assessment design considers which types of evidence would ideally be useful to reason about student learning and infer what students know and can do. Assessment is framed as the process of designing observational contexts (i.e., assessment tasks) that provide such evidence in service of some question, claim, or inference.

The ECD framework consists of five layers: domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery. Of these, the conceptual assessment framework is most relevant for discussion of evidentiary reasoning and assessment arguments (see Figure 2). The conceptual assessment framework (CAF) is a formal specification of the operational elements of an assessment, including construct definition, item design, scoring models, and statistical estimation of abilities. Though all of these elements exist implicitly in any assessment design, an ECD-based approach makes each component and its effect on the assessment system explicit through the three CAF sub-models: the student model, the evidence model, and the task model.

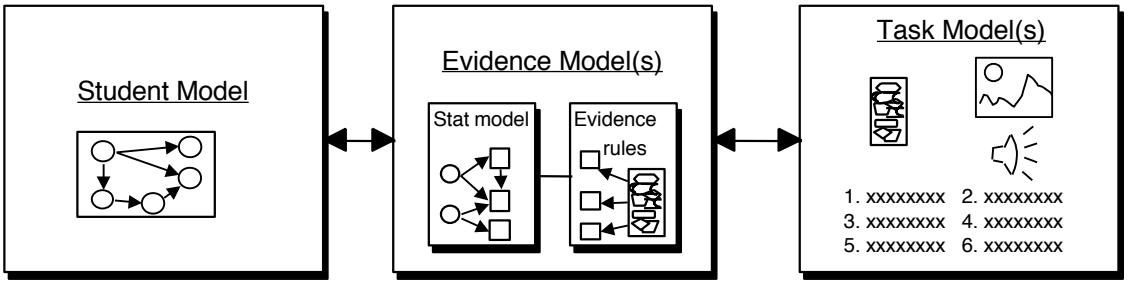

Figure 2. The conceptual assessment framework layer of Mislevy's ECD approach.

The *student model* defines our assessment goals in terms of the claims and inferences we wish to make as a result of our assessment. In traditional educational testing, the measurement target is some aspect of knowledge, skills, and abilities (KSA). It is common for educational assessment design to proceed from some theoretical conceptualization of a construct or domain to be measured – a construct definition. As compared to a construct definition, however, the student model is formulated as a more comprehensive model of our assessment goals. The student model includes not only a description of the KSAs we wish to measure, but incorporates our beliefs about the KSA's – their structure, development, and effect on behavior. The remainder of the assessment operates in service of the student model. It constitutes the claim of our evidentiary argument.

The additional two models, the *evidence model* and the *task model,* encompass the structure and elements of the evidentiary component of the argument. Figure 3 illustrates how data from these models could be structured into a Toulmin-like argument structure to support the claims of the student model. The *evidence model* of ECD describes salient features of students' observable behavior when interacting with the assessment tasks. This information is then used to update beliefs about student model variables through statistical models, also a component of the evidence model. When evidence is informative with respect to student model claims, our evidentiary argument is strengthened. Statistically speaking, when our observable indicators are highly reflective of our latent variables of interest, we can make accurate and valid claims. The key is to gain access to the best evidence to support our claims. That leads us to the final layer of the CAF, the task model.

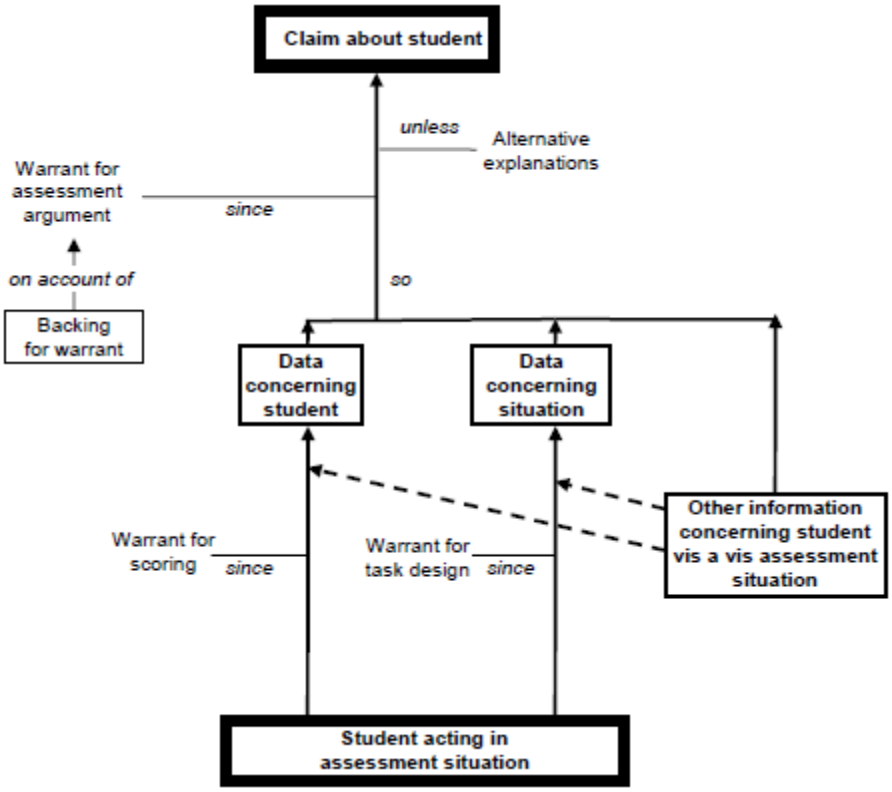

Figure 3. Illustration of assessment argument from Mislevy and Yin (2009).

The *task model*, as its name suggests, defines the characteristics of the assessment task, namely task/item features and the conditions under which the assessment is completed. The primary goal of the task model is to depict the environment in which a student will exhibit observable behaviors that correspond closely to the evidence models. Recall that in evidentiary arguments, the backing and warrant are critical for connecting the evidence to the claim. A carefully described task model should specify how each of the design choices (e.g., item format, available resources, time limits) relates to one or more pieces of evidence based on existing empirical knowledge or strong theoretical foundations. The strength of the overall assessment argument is enhanced through alignment of the task model and the evidence model.

The function of ECD, at its core, is to elucidate critical aspects of an assessment to make them more explicit, thereby improving the quality of the assessment argument and ultimately score interpretation and use. It does so by differentiating key subsystems of an assessment related to assessment design, implementation, and delivery. To summarize, Mislevy and Yin (2009) describe the role ECD as "explicating assessment as evidentiary argument brings out its

underlying structure, clarifies the roles of the observable elements and processes, and guides the construction of tasks. (p. 252). By deconstructing the larger complex assessment system into its component parts, more attention is given to some of the assumptions that are often made (with relatively little consideration) when designing and implementing assessments.

Existing Practices in Assessment as Evidential Argument

Arguably, the majority of educational assessments, particularly those designed for large-scale accountability assessment, were not designed from an evidential argument perspective. However, examples of evidential reasoning are evident in two less publicized forms of educational assessment: alternate assessments and psycho-educational assessment. These examples are used to highlight ideal practices of evidential reasoning that should be increasingly applied throughout educational assessment.

Alternate Assessments

The most obvious example of evidential reasoning in educational assessment, is applied to a surprisingly small percentage of students – the alternate assessment. Alternate assessments, reserved for the "1% population" of students, students with severe cognitive abilities, are used for accountability measurement for students who are not expected to perform at grade-level as a result of their disabilities. Unlike standardized tests for the general student population, which are fairly consistent in format and content across all states, alternate assessment practices vary greatly (See Schafer & Lissitz, 2009, for a comprehensive review). However, based on trends evident in policies and practices in many states, alternate assessments are generally better models of evidentiary reasoning for student assessment than that implemented for the general student population (Elliott & Roach, 2007; Ysseldyke & Olsen, 1997). Strangely, this practice stems from the lack of appropriate "standardized" measures to use with the 1% student population. As a proxy for the single standardized test used with the general education population, alternate assessments draw from behavioral observations, performance assessments, rating scales, checklists, portfolios, and sometimes test performance. Though the nature of each of these indicators is unique, each requires collection of evidence samples (e.g., classroom work products, videotapes, interviews, structure observations, students' responses to on-demand tasks) presumed to characterize students' knowledge and skills of interest. The evidence is then scored

to yield data and that score is interpreted, typically relative to performance standards. Additionally, these assessments almost without exception consider auxiliary information about the student in terms of past performance, knowledge of specific abilities (or disabilities), and other background information that could affect assessment claims and evidence interpretation. Elliott and Roach (2007) illustrated the underlying logic and implementation of alternate assessments as a multi-stage evidence based model (see Figure 4). Surprising similarities to components of Toulmin's evidentiary argument structure are represented. The proficiency level decisions are the claims of the argument. The evidence samples and scores are the data. The correspondence between these and the grade level content standards, as well as the use of multiple raters provide backing and warrants.

Figure 4. Common framework across various forms of alternate assessment (adapted from Elliott & Roach, 2007)

Unfortunately, though alternate assessments implement the general framework and principles of evidentiary reasoning, they have significant weaknesses and are of questionable use and quality. Much of the criticism of alternate assessment stems from a lack of backing to support the use of various evidence sources leading to questionable reliability and validity of the score interpretations. The challenges in alternate assessment are most likely to occur during alignment, scoring, and standard setting (Elliott & Roach, 2007). Further, the validity arguments are often lacking in terms of empirical evidence associating the observed scores with the targeted skills and knowledge (Goldstein & Behuniak, 2011). These problems arise for various reasons stemming from the nature of alternate assessments and their targeted population. As compared to general education assessments, the small samples, heterogeneous student population, inconsistent opportunities-to-learn, and variability across alternate assessment practices make it difficult to estimate any of the psychometric properties and meaning of the scores derived from these tests (Rodruigez, 2009). Consequently, efforts are underway to develop measurement models that can

help support the development, use, and interpretation of alternate assessments. Specific interest lies in how to use these assessments to measure growth and status of students with disabilities for accountability purposes. Still, the general approach to alternate assessments which considers broader evidence sources and more highly contextualized information about student behavior is instructive. Whereas the alternate assessments draw from multiple sources of evidence useful for backing assessment claims, the traditional end-of-year general education approach relies solely on a single evidence source. General education assessments would be improved by adopting similar evidentiary models and practices.

Psycho-educational Assessment

Psycho-educational assessment, like that performed by school psychologists and counselors, also offers an ideal example of multi-source evidence-based reasoning. When parents or teachers refer children for assessment, the goal is to identify the underlying causal source of a student's classroom or home behavior. Typically, the parent or teacher has noticed that a child is not performing in a "typical" manner, either cognitively, emotionally, or behaviorally. The goal of the assessment is to diagnose the underlying cause of the behavior and prescribe a prescriptive course of action. Wodrich and Schmitt (2006) provide a framework to proceed through an assessment in the most efficient, reliable, and valid manner so as to make the strongest claim about a child's instructional needs (see Figure 5). Though not presented here, within each box, a series of data-sources are used to make the yes-no determination. At each step, the probabilities of various claims (i.e., "the child has a non-verbal learning disability" or "the child has ADHD") change, based on the evidence that is provided. At the end of the assessment process, the probability of one of the claims should be higher than any other claim, making it the most likely diagnosis.

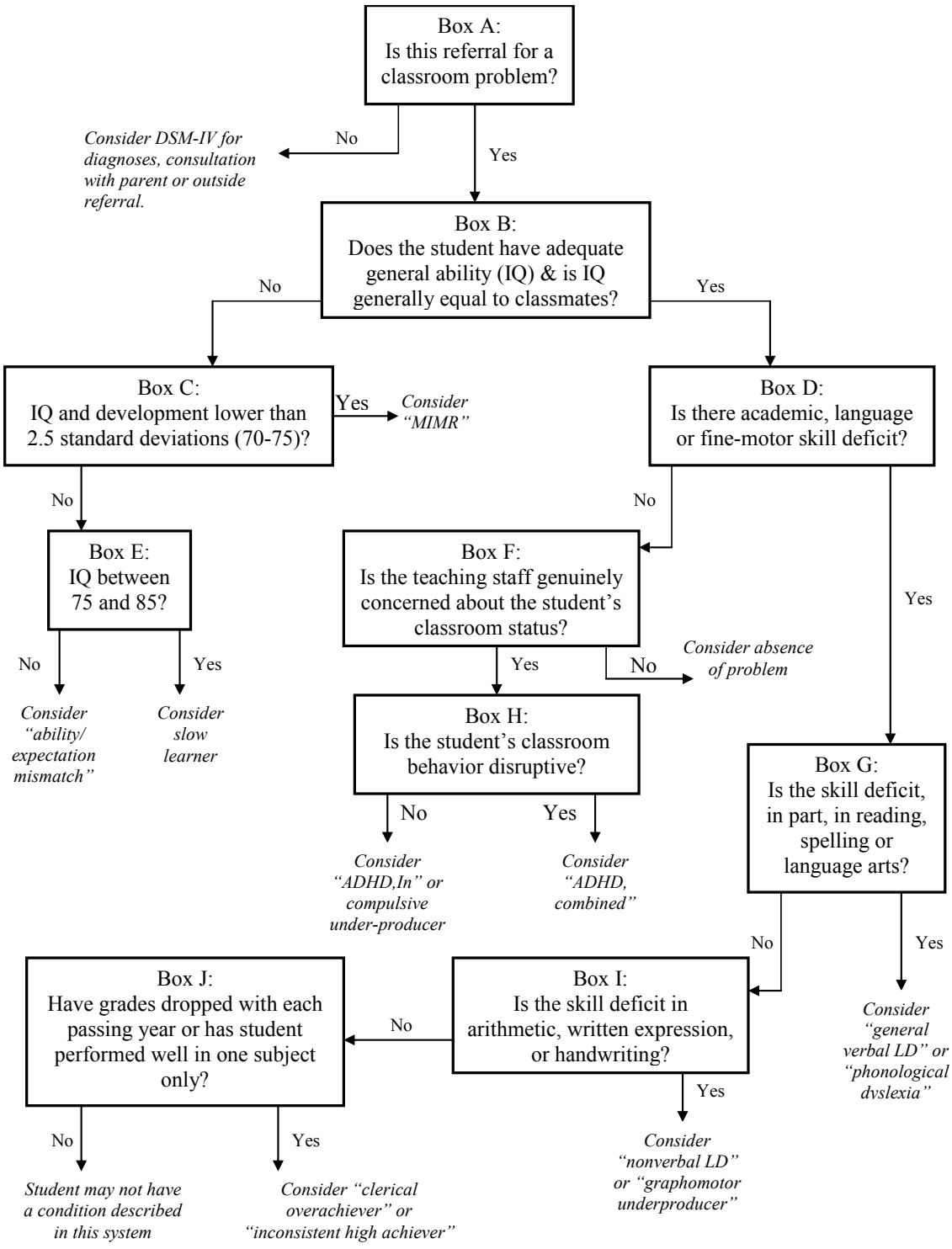

Figure 5. Wodrich & Schmitt's (2006) decision framework for psycho-educational assessment.

Several aspects of Wodrich and Schmitt's (2006) system are particularly important for consideration in educational assessment. First, the sources of evidence are quite varied. In some

cases, evidence is derived from third-party reports about behavior; in other cases, decisions are made based on scores from standardized achievement and ability tests. Second, the nature of the evidence varies from quantitative (e.g., standardized scale scores) to qualitative (e.g. judgments of teacher concerns). Finally, aside from scores on the standardized tests, the evidence is obtained from the actual contexts in which the students engage on a daily basis – the contexts in which we are interested in making claims. Kaufman (1979) described the process of appropriately integrating multi-source evidence as *intelligent testing*, focusing on the merger between measurement science (based on psychological theory) and clinical assessment. Though the two fields were quite distinct through most of the 20th century, they are now inseparable for purposes of educational assessment. Each offers backing and warrants for distinct sets of evidence that contribute to the assessment and diagnostic process. However, consideration of one, without qualifications and consideration of the other, is likely to lead to poorly supported claims and incorrect diagnoses. As illustrated by Wodrich and Schmitt's decision tree, the manner and order in which the evidence is collected and integrated is critical to arrive at a valid assessment claim.

Just as with the case of the alternate assessments, the goal of the psycho-educational assessment is to support decisions regarding students' instructional experience and curriculum, which exist within the classroom. The assessment process thus relies heavily on information about how the student currently behaves in the classroom. As part of the decision tree, it is critical that evidence about the classroom behavior of the students be differentiated from and compared to more clinical measures that are "context-free." Although the "skills" required by both classroom activities and standardized tests may be the same, the classroom-based measures contextualize the assessment tasks in such a way as to fundamentally change the meaning of the scores. Whereas the standardized tests measure more "pure abilities", the classroom measures (including teacher observations and student work products) reflect student motivation, engagement, time management, strategies, and self-regulation, simultaneously. These varied data sources, including those gathered in the authentic environment of interest are needed to parse among competing explanations for the student's referring problem. The use of both standardized measures and classroom-based data is necessary to select among the possible diagnoses and prescribed intervention.

Though the goal of educational assessment for accountability is not typically diagnosis as in the example given here, two principles of evidentiary argumentation for diagnostic testing should transfer well to accountability assessment – multiple, varied data sources and evidence from appropriate, rich contexts. Every assessment claim is a form of diagnosis. The parallel between medical or psychological diagnosis and educational testing is more compelling when we consider the increasing interest in diagnostic score interpretations from traditional skills-based tests (Levy, 2011a). No Child Left Behind (NCLB, 2001) legislation includes provisions for increased use of educational test scores for the purpose of instructional design and remediation. The goal is to increase the instructional utility of test scores. In order to be successful in tying assessment results to student learning and classroom instruction, the assessment argument must incorporate the learning context, or the conclusions are apt to be invalid for any real world purpose. Otherwise, we risk collecting data that can only be used to make claims about abilities in isolation, which is of little use to the educational process. I turn now to consideration of how evidential reasoning principles could be incorporated into large-scale assessment of general education populations for accountability purposes.

Implications for Assessment Design and Development Practice

Shifting our perspective on assessment to that of an evidential argument is more than a theoretical change; it has profound implications for educational assessment design and development practice, as suggested in our discussion of the ECD approach. As argued earlier, the strength of any argument lies primarily in the quality of the evidence (i.e., data) and the warrant regarding its relationship to the claim. The quality of the evidence is judged relative to the specific claim; evidence that is persuasive for one claim may be useful for another. In order to build the strongest argument, one should work backwards from the claim by addressing the question of "what evidence would be persuasive of the claim I want to make"? Then ask "what situation will give rise to such evidence?" The future of educational assessment will be determined by our answers to these questions, answers that may look different from those today for three reasons: a) changes in the nature of the claims we want to make about students, b) availability of new data sources that could inform the existing argument, and c) new analytic tools to translate data into usable evidence that supports or refutes the claim.

New Assessment Claims

A well-formulated evidentiary argument includes only evidence that is relevant to its claim. Data that is informative evidence for one claim may be irrelevant for another. As evidential arguments, our assessments should be designed to carefully elicit that data which serves our evidentiary needs. We must therefore carefully consider whether the claims we want to make from our assessments are in fact those that are supported by current assessments. That is, do our current educational assessments provide evidence about those "things" we want to know about students?

New constructs. An informal review of state and national educational assessment systems for accountability reveals a strong emphasis on what some call "basic skills". Though this trend was likely motivated by a desire to improve individuals learning, skill, and overall educational opportunity, the result has been a narrowing of curriculum, a de-emphasis of elective curricula (e.g., performing arts, foreign language), and neglect of critical higher-order reasoning skills that are critical for success in today's society (Crocco & Costigan, 2007). It is unclear whether the use of high-stakes assessments has driven this change or is merely following the curricular shift. Regardless of this distinction, clearly our assessments are quite narrowly focused in terms of a small set of constructs and their operational definitions. Unfortunately, despite attempts to improve student learning through the "basic skills" curriculum and assessment approach, international trends continue to show that U.S. students lag behind those of other countries, countries whose curriculum is designed to emphasize critical thinking, reasoning, and problem-solving skills (Chen, Gorin, Thompson, & Tatsuoka, 2008; Corter & Tatsuoka, 2004). The ultimate effect is that employers are increasingly dissatisfied with graduating high school and college students' abilities to deal with real-world problem solving and other critical tasks in the workplace (Casner-Lotto, 2006).

21st Century Skills assessment. As the workplace and global economy change ever more rapidly, so do the specific sets of knowledge and skills needed for individuals and institutions to be successful. To capture the unique set of skills currently "in-demand", the term *21st century skills* was coined to distinguish the skills and knowledge needed for success in today's workplace (Silva, 2008). Educators and the educational system are pressed to adapt to these changing needs by re-focusing curriculum on these 21st century skills rather than more "basic skills" (Gee, 2010). As opposed to the traditional curriculum that focuses on highly de-contextualized component

skills, alternative models of learning and education focus curriculum and assessments on students' ability to apply their "affective, cognitive, and situative processes to solving the problems of living" – whatever those problems may be and whatever processes those require. Gordon (2007) called this *intellective competence*. Intellective competence is "a characteristic way of adapting, appreciating, knowing, and understanding the phenomena of human experience," as well as "the quality with which these mental processes are applied in one's engagement with common, novel, and specialized problems."

Educational assessments must reflect educational goals and instructional objectives. With new claims come new set of evidence, and new tasks to generate the supporting data. Perhaps the most significant effect of claims about intellective competence for the future of assessment resides in the structure of the construct. Unlike most "basic-skills," processes encompassed by intellective competence are multi-dimensional, cross-contextual, and cross-disciplinary, at their core. They emphasize individuals' ability to use their KSA's in service of some goal – a goal larger than answering a test question correctly. Intellective competence captures individuals' ability to monitor one's progress, work collaboratively with others, and engage in situations that are both social and cognitive in nature. Unlike indicators of traditional skills, intellective competence assessment should provide information useful for describing and predicting real-world behavior. Though not referring specifically to intellective competence, the North Central Regional Education Laboratory asserted that the "assessment of student achievement is changing, largely because today's students face a world that will demand new knowledge and abilities" (Bond, 2012). Where our existing assessments fall short in predicting outcomes[1] such as academic success and college graduation rates, measures of intellective competence should provide strong evidence to support claims regarding students' progress towards becoming productive members of an economically solvent society.

Assessing process for diagnostic assessment. Assessment claims are more than just the construct targeted by an assessment; they speak directly to the intended use of the assessment scores. As mentioned earlier, there is a trend in educational assessment to use test scores for diagnostic purposes, which support prescriptive instructional design to enhance individual and group learning (Bennett, 2010, 2011; FAST SCASS, 2008; NRC, 2001). Whereas traditional

[1] Numerous studies have shown the moderate correlations between standardized admissions test scores and achievement indicators, including GPA and graduation rates. See Zwick (2007) for a review of the use of standardized admissions tests in higher education.

status measurement allows policy makers and educators to take stock of individuals' and groups' current level of KSAs, diagnostic assessment, provides a basis for making claims about *why* individuals are at a particular level and *how* they can improve their levels. The evidentiary grounding needed to support diagnostic assessment claims is perhaps more demanding than for status assessment. Evidence must be available from student responses that isolates weaknesses or inconsistencies in knowledge and provides a fuller picture of student abilities. Valid claims about how instruction can be improved require evidence of how students' cognitively process content and context and the process by which they arrive at correct or, more importantly, incorrect answers (Wylie, 2012). Determining whether an incorrect response results from a) misconceptions about domain knowledge, or b) poor monitoring and use of available information, is a critical distinction to be made for diagnosis and instructional planning.

Good diagnostic assessments have been distinguished in terms of their *penetration* (Cross & Paris, 1987). Penetration is defined as the resulting psychological information obtained from the test's scores, including information about concepts, knowledge representations, and cognitive processing. Penetrating tests provide evidence of individual knowledge and processing by providing opportunities to observe the process of student responses and increase the amount of information available from student answers. To this end, diagnostic assessments must generate evidence of student processes including strategy use, processing speed, attentional control, response selection, and self-regulatory processes. Assessment tasks must be designed to elicit behaviors that provide the necessary evidential data – response times, student-log data, navigational patterns, eye-movements, all of which (Gorin, 2007).The use of innovative tasks designs to capture evidence of process and status is discussed later in this chapter.

New theoretical models. While a change to the list of assessment constructs has emerged more for practical and economic reasons, a less obvious change has transpired at the theoretical level with respect to our beliefs about learning, which could have equally significant impact on the future of educational assessment design. In 2001, the National Research Council (NRC, 2001) highlighted four perspectives on the nature of the human mind: the differential perspective, the behaviorist perspective, the cognitive perspective, and the situative perspective. All four perspectives define principles by which human cognition and behavior can be described and predicted, though each has a slightly different focus. The differential perspective, which grows out of some of the earlier work on individuals differences research and abilities testing of

the early 1900s, focuses on differences in what people know and their potential to learn. The behaviorist perspective focuses on acquisition of knowledge in terms of stimulus-response associations as the building blocks of learning. The cognitive perspective also focuses on structure of knowledge, but broadens the scope of possible structures beyond the behaviorist model. As compared to all three of these perspectives, the situative or sociocultural perspective considers the individual within a context, rather than in isolation. The focus of this perspective is on characterizing individuals' real-world cognition and behavior as they interact with their own environment.

The focus on interactions between individuals' cognition and the situative context only serves to heighten the importance of task design in the assessment argument. Whereas trait based models of behavior reduce the set of explanatory variables to a single or set of relatively stable characteristics of the individual, interactionist models simultaneously consider individual, situative, and cultural influences (Gee, 2007, 2010). In order to build an assessment argument based on a situative model of learning, tasks must be used that generate evidence about individuals' interactions with the assessment context. This is a much broader and demanding evidentiary requirement than more traditional cognitive models of learning that focused solely on evidence regarding the "internal" cognition of the individual. In order to capture the necessary evidence to support claims about situated learning and cognition, new evidence sources with appropriately designed context must be developed.

New Evidence Sources

Traditionally, educational assessments have relied heavily on scored responses to paper-pencil group-administered tests. In the early to mid 1900s, when large numbers of military recruits needed to be tested, the paper-pencil format was the only feasible option for practical reasons. Sadly, with almost a century having passed, the changes that have occurred are, for the most part, superficial. Transitions from paper-pencil to computer administration of multiple-choice tests merely changed the delivery mode of the identical items, rather than a change to the nature of the tasks and the measured constructs. Some notable exceptions to the relatively stagnant assessment practices of the 20th century are the implementation of computerized adaptive testing based on item response theory models, and the use of construct response item formats, most commonly short answer and essay items on large-scale tests. Still, until recently,

what has remained relatively unchanged since the origins of large-scale educational testing is the general approach – sit a student down at a test for a single testing session for a brief period of time, collect responses to a set of relatively decontextualized items, fit a unidimensional model of proficiency, and describe students' ability in terms of performance relative to one another or to some content-related standard. In terms of building an evidential argument, this "drop-in-from-the-sky"[2] approach yields a relatively weak evidentiary foundation for making claims about individuals' ability to reason and function in the world. More likely, evidence from current tests support claims about a student's ability to use knowledge of a series of basic skills, knowledge of test taking strategies, and motivation to score well on a test (though by reporting a single score, as is typical, we cannot even distinguish the effects of each of these factors on student performance). Claims about student success in higher-education institutions and ultimately in the workplace are only weakly supported by this evidence.

This argues for the need of new sources of evidence more closely tied to the types of claims employers and policy makers want to make (Silva, 2008). Recent and future assessments that widen the assessment frame to include multiple evidence sources, from varied data, across numerous contexts should provide a more robust argument to support our claims about students' learning and their ability to navigate contemporary society.

While the technological revolution of the 21st century has undeniably changed the types of claims we wish to make about students' learning and abilities, technology has had an arguably more dramatic change on the way in which we collect evidence to support those claims. As shown in the illustration of the assessment argument, several types of data provide evidence, connecting data to claims: data concerning students, data concerning the assessment situation, and data concerning students vis-à-vis the assessment situation. Technological innovations have expanded our capabilities to capture student behaviors relevant to all three types of evidence. As summarized by Shute (2011)

> *…new technologies allow us to embed assessments into the learning process; extract ongoing, multifaceted information (evidence) from a learner; … we can support learning by using automated scoring and machine-based reasoning techniques to infer things that would be too hard for humans (e.g., estimating*

[2] Mislevy has used the term "drop-in-from-the-sky" to refer to traditional testing practices in which students stop their typical learning activities in order to complete the assessment tasks on a pre-specified day and time.

competency levels across a network of skills, addressing what the person knows and can do, and to what degree).

Item types. Some of the most exciting assessment developments in recent years are those involving innovative assessment tasks and item types, including the use of scenario-based items and simulations. In recent decades the use of performance assessments, specifically constructed-response items in lieu of forced-choice items (i.e., multiple-choice items) increased significantly. Performance assessments are presumed to be more sensitive to higher-level thinking skills and more "authentic" in terms of the measured construct than traditional "artificial" item types. As a result, most high-stakes assessment systems now include some form of constructed response item as part of their tests. This shift has been facilitated by improved technologies that support automated scoring of constructed responses, including automated essay scoring systems such as *eRater*© and *cRater*© (Shermis & Burstein, 2003).

Still, critics of current educational assessments would argue that most operational constructed-response assessment tasks are only a slight improvement over their predecessors, particularly if we consider the recent interest in situated models of cognition and learning. They still suffer from the limitation of more objective item types in that that they rely heavily on a limited set of behavioral observations gathered in an artificial testing context. Both the traditional forced-choice and the brief constructed-response tasks lack appropriately situated contexts to generate evidence of how students interact with and reason in their natural environments.

New assessment tasks should incorporate items that require processing consistent with contemporary models of student learning and cognition (Gee, 2007). Specifically, items and tasks that develop rich contexts within which individuals must reason and respond, similar to real-world cognition are of interest. Several new task formats offer promising opportunities: scenario-based tasks, simulations-based assessments, and educational games. Scenario-based tasks embed traditional test questions into an artificial context typically presented through text or video. The actual test items will usually be presented as a set in multiple-choice or constructed-response format. However, the nature of the question will be tied to the context presented in the scenario. Scenario-based assessment (SBA) has been used consistently on assessment in various professional disciplines, most predominantly in medical fields (Lurie, 2011).

A variation on scenario-based assessment is the use of simulation-based assessment tasks. In simulations, the scenario is created to mimic as closely as possible, all the components and

functioning of the real world. As compared to scenario-based tasks, however, the context is presented more realistically in simulations, typically allowing examinees to interact with elements of the task in a manner analogous to the real-world environment. One of the most successful simulation-based assessments is Cisco's *Networking Performance Skill System (NetPASS:* Behrens, Mislevy, Bauer, Williamson, & Levy, 2004). NetPASS is designed primarily to measure networking proficiency, but also to provide feedback to e-learning students and instructors as well as use for certification decisions. The tasks mimic actual Cisco networking equipment with specific design constraints similar to real world "scenarios" of network troubleshooting, design, and implementation (Williamson, Bauer, Steinberg, Mislevy, Behrens, & DeMark, 2004). Examinees must design, implement, and/or troubleshoot networks based on typical network failures, configuration requirements, and design constraints. To solve the problem, movable icons are provided representing all elements of a network allowing examinees to solve the simulated problem just as they would a real network in an office building. By embedding the assessment in a specific simulated context, alignment between the evidence and the targeted assessment claims is enhanced. The claims are about performance in a networking environment and the evidence is collected in a networking environment.

While both scenario-based and simulation-based tasks embed the assessment in context aligned with the assessment claims (e.g., a networking environment for assessing networking skills), some have suggested that the use of unrealistic environments that do not resemble any real-world context may be even more powerful assessment tools. I am referring to the use of educational games as assessments (Gee, 2007, 2010; Shute, 2011). Educational games for instruction are expected to enhance learning by offering a structured learning space in which the complexity, sequencing, and frequency of curricular objectives can be controlled (Gee, 2010). Game interfaces that allow for the use of imaginary "worlds" that defy reality, can increase interest and engagement for many school-aged children, who play similar games as recreational activities by choice – often for hours at a time. Further, unlike existing assessment contexts, games are played out within a social context, one in which actors must learn the rules of engagement and community standard operating procedures in order to garner assistance from others and ultimately be successful in the game. As our desired assessment claims expand to include inferences about individuals' intellective competencies, including their ability to

strategically navigate real-world social contexts, games as assessments may be our first opportunity to collect the necessary evidence to support our argument.

Novel data sources. As just described, the use of technology-enhanced assessments can strengthen our assessment arguments by creating a task environment that includes more of the skills and processes we are interested in measuring. A useful byproduct that is equally important for building our evidentiary arguments in future assessment is the on-line data that can be captured and scored. Unlike paper-pencil or other non-computer administered tasks, computer-based assessments can leverage the technology to capture a variety of data on student interaction with the assessment tasks. The simplest example is that of examinee response time. A rich history exists to support the use of response times as indicators of cognitive abilities (Lee & Chen, 2011; Schnipke & Scrams, 2002). Response times have traditionally provided evidence of processing speed, as opposed to or in relation to processing accuracy. More recently, interest has shifted to their use as evidence of additional constructs including student motivation and engagement (Wise & DeMars, 2006). Unexpectedly short response times can indicate rapid-guessing or low engagement, both of which could be alter the meaning of test scores and associated inferences (De Mars, 2007). Within the evidentiary argument, response times could be used as qualifiers that mediate the relationship between evidence from item response accuracy and the assessment claims.

A more complex data source, student log-data, can also be easily recorded as students navigate through a computer-administered test. Whereas response times only tell us about the duration of student interaction with the assessment, log-data provides information about exactly what students were doing with the task at a given time. Key strokes, mouse clicks, scrolling. All student-computer interactions can be captured and examined as potential evidence regarding student learning and knowledge. Returning to the NetPASS simulation previously described, log-data is captured and scored as part of the assessment. Rather than scoring only the accuracy of the final network assembly, the logs of all computer workstation commands are collected as evidence of claims regarding the completeness and the correctness of procedures while solving the problem. In addition to certification decisions and overall scores, diagnoses of specific problems are made, by comparing process and outcome of student logs to previously identified processing patterns associated with known weaknesses or misconceptions. Of course, the challenge to this approach is to establish some criteria by which examinee's log-command

pattern can be compared. Expert task analysis and other methods of cognitive modeling are useful tools. Little research exists on how log-data can be analyzed and scored, but the possibilities offer an exciting new evidence source for our assessment arguments.

Well outside the scope of traditional assessment data, some research has begun to examine the utility of psycho-sensory data to measure student engagement and attention when learning and being assessed (Sanchez, et al., 2011). Physiological measures include pupil dilation, eye-movement and fixations, and electromagnetic-brain activity. Though relatively rare, eye-tracking data has been used in the context of educational assessment as part of the validity argument. Several researchers have used experimental eye-tracking studies to provide evidence of the cognitive processes and attentional resources allocated by students when solving problems (Gorin, 2006; Ivie et al, 2004). Physiological data, such as that collected from eye-tracking, further has the potential to be more directly part of the assessment argument, not only as part of the validity argument. Considerable research on the backing and warrants for these evidence sources is still needed before operational use is feasible.

New Analytic Tools: Translating Data into Evidence

Ultimately, in order to make inferences about our claims using observed evidence, the data must be translated into interpretable form. Quantitative psychometric approaches, ranging from classical true score theory to item response theory, offer a variety of methods for converting individuals' behaviors into estimated ability levels. In the traditional assessment paradigm the focus is on transforming scored item responses into latent trait estimates. Within the perspective, the majority of research efforts to improve our psychometric capabilities have focused on several key issues for estimating latent traits, namely advanced methods for dimensionality assessment (e.g., Levy, 2011b), multi-dimensional and higher-order latent trait models (e.g., de la Torre & Douglas, 2004), and "diagnostic" and explanatory models (e.g., de Boeck & Wilson, 2004; Rupp, Templin, & Henson, 2010). The computational sophistication of the models has grown dramatically since the early unidimensional latent trait models at the inception of modern measurement theory (Lord & Novick, 1968).

If assessments are to provide usable evidence to support the more complex interactionist claims advanced earlier in this chapter, our analytic tools must also adapt. At the very least, unidimensional models that assume a single underlying latent trait affecting task performance are

overly simple for highly contextualized tasks appropriate to measure intellective competency. Multi-dimensional IRT models, for example, allow simultaneous consideration of multiple abilities that interact to produce a response. However, these models still typically only model a single piece of observed data, an item-level score.

Assessment as evidential reasoning may require a greater departure from our traditional modeling approaches than a mere increase in dimensional space. Perhaps a more philosophical shift in our thinking about our modeling goals is needed. From a statistical perspective, complex evidentiary arguments can be viewed as a set of probabilistic relations connecting our observed data to our claims (Levy, 2011a; Mislevy & Levy, 2007). Advances in computational capabilities and statistical modeling techniques now permit mathematical estimation of assessment arguments in terms of probabilistic models of the argument components (i.e., data, claims, warrants, qualifiers, etc.). Though not originally conceived within a measurement or assessment context, statistical modeling techniques, such as Bayesian inference networks (BINs), are useful for incorporating multiple and heterogeneous evidence into reasoning or argumentation (Schum, 2001). When including the observed data as evidence and students' ability and trait levels as our claims, model parameter estimates and fit indices provide an empirical test of our assessment argument.

Figure 7 shows a fragment of a conditional probability model corresponding to elements of the evidentiary argument underlying the NetPASS assessment. The arrows between each element in the model represent conditional relationships among the student model claims, the observable evidence, and task-specific features to be estimated via some statistical model. The strength of the conditional relationships between observable variables from the tasks and the student proficiency variables and the fit of the overall model provides evidence in support of the hypothesized underlying cognitive processes. Further, once stable estimates of the paths are estimated, observed data from any examinee can be used to estimate the multiple proficiencies targeted by the test (e.g., Declarative Knowledge, Troubleshooting Procedures, and Network Modeling). Future assessments based on evidentiary reasoning principles will require analytic tools like BINs, which support larger amounts of more complex data into a single probabilistic model (Conati, Geriner, & Van Lehn, 2002; Mislevy, 1994).

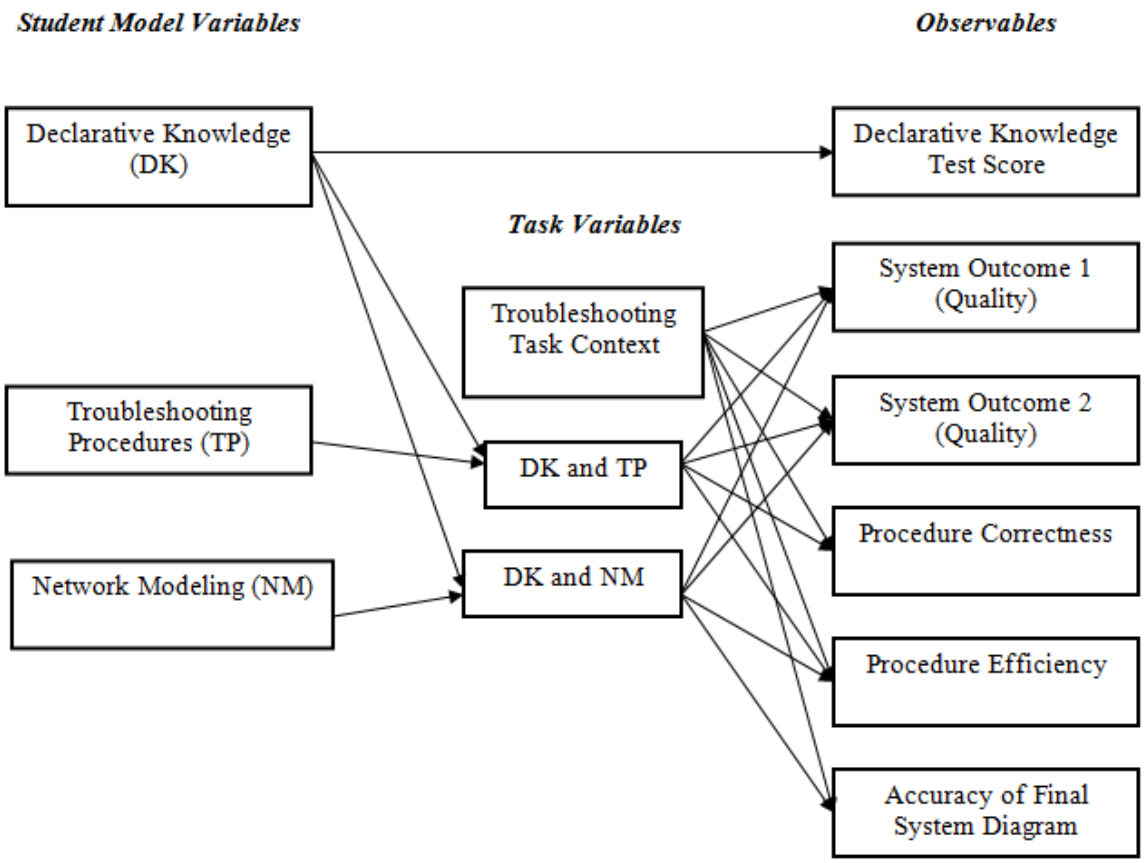

Figure 7. Fragment of conditional probability model for estimation with BINs for NetPASS Troubleshooting Task (adapted from Mislevy, 2004).

Conclusion

At its core, an evidentiary argument is defined by the claims it seeks to substantiate or refute. Perhaps the most common criticism of educational tests as the entirety of educational assessment is the narrow focus on the types of claims they can support. The goal of the educational system arguably is not to incrementally increase students' ability to answer isolated questions correctly. Rather, the goal is to capture and understand individuals' capability to interact with one another and their environment in more strategic, adaptive, and successful ways. Educational assessments must reflect this goal. Recent developments in technology, cognitive and learning theory, and measurement and psychometrics have each had unique impact on modern educational assessment. However, assessment as evidentiary reasoning about the claims that interest us in the 21st Century and beyond requires a more integrated consideration of these related fields. Educational assessment models should parallel our complex cognitive,

sociocultural models of learning. The psychometric models should handle multiple types of data and consider parameters that reflect individual and situational factors. The view of assessment as a one-hour, one-day, or one-week scheduled effort must be eradicated. The dynamic processes that should be targeted by educational assessment, if appropriately captured, requires evidence that keeps up with the real-time changes occurring within and around students as they interact with the world. If we are successful in our efforts, then the future of assessment should look more like every-day real-world interactions than our typical notion of an educational test.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1999). Standards on Educational and Psychological Testing.

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning: A preliminary theory of action for summative and formative assessment. Measurement: Interdisciplinary Research and Perspectives, 8, 70-91.

Bennett, R. E. (2011). Formative assessment: A critical review. Assessment in Education: Principles, Policy and Practice 18, 5-25.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & 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, 295-302.

Casner-Lotto, J. (2006). Are they really ready to work?: Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce. U.S.: Conference Board.

Chen, Y. H., Gorin, J. S., Thompson, M. S., & Tatsouka, K. K. (2008). Cross-cultural validity of the TIMSS-1999 Mathematics Test: Verification of a cognitive model. International Journal of Testing, 8(3), 251-271.

Conati, C., Geriner, A., & Van Lehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. User Modeling & User-Adapted Interaction, 72(4), 371-417.

Corter, J. & Tatsuoka, K. K. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. American Educational Research Journal, 41(4), 901-926.

Crocco, M. S., & Costigan, A. T. (2007). The narrowing of curriculum and pedagogy in the age of accountability: Urban educators speak out. Urban Education (42), 512-535.

Cross, D. R., & Paris, S. G. (1987). Assessment of reading comprehension: Matching test purposes and test properties. Educational Psychologist, 22(3&4), 313 – 332.

Cronbach, L. J. (1989). Construct validation after thirty years. In L. J. Cronbach (Ed.) Intelligence: Measurement, theory, and public policy: Proceedings of a symposium in honor of Lloyd G. Humphreys (pp. 147-171). Champaign, IL: University of Illinois Press.

de Boeck, P., & Wilson, M. (Eds.). (2004). Explanatory item response models: A generalized linear and nonlinear approach. New York: Springer.

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

De Mars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. Educational measurement, 12(1), 23-45.

Elliott, S. N. & Roach, A. T. (2007). Alternate assessments of students with significant disabilities: Alternative approaches, common technical challenges. Applied Measurement in Education, 20(3), 301-333.

Formative Assessment for Students and Teachers (FAST) State Collaborative on Assessment and Student Standards (SCASS). (2008, October). Attributes of effective formative assessment. Paper prepared for the Formative Assessment for Teachers and Students State Collaborative on Assessment and Student Standards of the Council of Chief State School Officers. Washington, DC: Council of Chief State School Officers.

Gee, J. P. (2007). Reflections on assessment from a sociocultural-situated perspective. In P. A. Moss (Ed.) Evidence and decision making (pp. 362-375). Blackwell Publishing.

Gee, J. P. (2010). Human Action and social groups as the natural home of assessment: Thoughts on 21st century learning and assessment. In V. J. Shute & B. J. Becker (Eds.) Innovative assessment for the 21st century: Supporting educational needs (pp. 13-39). Springer Science.

Goldstein, J. & Behuniak, P. (2011). Assumptions in alternate assessment: An argument-based approach to validation. Assessment For Effective Intervention, 36(3), 179-191

Gordon, E. W (2007). Intellective Competence. Voices in Urban Education: Towards Proficiency, 14.

Gorin, J. S. (2006). Using alternative data sources to inform item difficulty modeling. Paper presented at the 2006 Annual Meeting of the National Council on Educational Measurement.

Gorin, J. S. (2007). Test Construction and Diagnostic Testing. In J. P. Leighton & M. J. Gierl, Eds. Cognitive Diagnostic Assessment in Education: Theory and Practice (pp. 173 - 204). Cambridge University Press.

Ivie, J. L., Kupzyk, K. A. & Embreston, S. E. (2004). Final report of Cognitive Components Study - Predicting strategies for solving multiple-choice quantitative reasoning items: An eyetracker study. Princeton, NJ: Educational Testing Services and Lawrence, KS: University of Kansas.

Kane, M. (1992). An argument-based approach to validation. Psychological Bulletin, 112, 527-535.

Kaufman, A. S. (1979). Intelligent testing with the WISC-R. New York: Wiley.

Lee, Y-H., Chen, H. (2011). A review of recent response-time analyses in educational testing. Psychological Test and Assessment Modeling, 53, 359-379.

Levy, R. (2011a). Evidentiary reasoning in diagnostic classification models. Measurement: Interdisciplinary Research and Perspectives, 7(1), 36-41.

Levy, R. (2011b). Posterior predictive model checking for conjunctive multidimensionality in item response theory. Journal of Educational and Behavioral Statistics, 36(5), 672-694.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test score. Reading, MA: Addison-Welsley Publishing Inc.

Lurie, S. (2011). Towards greater clarity in the role of ambiguity in clinical reasoning. Medical Education, 45(4), 326-328.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) Educational Measurement (3rd ed.) (pp.13-103). New York: American Council on Education/Macmillan Publishing.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. Psychometrika, 59, 439-483

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.) Educational Measurement (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.

Mislevy, R. J. (2012). Four metaphors we need to understand assessment. Princeton, NJ: The Gordon Commission on the Future of Assessment in Education.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. Educational Measurement: Issues and Practice, 25(4), 6-20.

Mislevy, R. J. & Levy, R. (2007). Bayesian psychometric modeling from an evidence-centered design perspective. In C. R. Rao and S. Sinharay (Eds.), Handbook of statistics, Volume 26 (pp. 839-865). North-Holland: Elsevier.

Mislevy, R. J., & Yin, C. (2009). If Language Is a Complex Adaptive System, What Is Language Assessment. Language Learning, 59(1), 249-267.

National Research Council (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.

Bond, L. A. (1995). Rethinking assessment and its role in supporting educational reform. North Central Regional Educational Laboratory. Retrieved September 10, 2012 from the World Wide Web: http://www.ncrel.org/sdrs/areas/issues/methods/assment/as700.htm.

Rodriguez, M. C. (2009). Psychometric considerations for alternate assessment based on modified academic achievement standards. Peabody Journal of Education, 84, 595-602.

Rupp, A., Templin, J., & Henson, R. (2010). Diagnostic Measurement: Theory, Methods, and Applications. New York, NY: Guilford Press.

Sanchez, J. G., Christopherson, R., Echeagaray, M. E. C., Gibson, D. C., Atkinson, R. K., & Burleson, W. (2011). How to Do Multimodal Detection of Affective States? ICALT, 654 – 655

Schafer, W. D., & Lissitz, R. W. (2009). Alternate Assessments Based on Alternate Achievement Standards: Policy, Practice, and Potential. Brookes Publishing Company.

Schnipke, D. L. & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenze, J. J. Fremer, & W. Ward (Eds.), Computer-based testing: Building the foundation for future assessments (pp.237-266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schum, D. A. (2001). The Evidential Foundations of Probabilistic Reasoning. Evanston, IL: Northwestern University Press.

Scott, I. A. (2009) Errors in clinical reasoning: causes and remedial strategies. British Medical Journal, 339(7711-7719).

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

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), The social science of serious games: Theories and applications (pp. 295-321). Mahwah, NJ: Routledge, Taylor and Francis.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J.D. Fletcher (Eds.) Computer games and instruction. (pp 503-524). Charlotte, NC: Information Age Publishing.

Silva, E. (2008). Measuring skills for the 21st century. Washington, D.C.: Education Sector Reports.

Toulmin, S. E. (1958). The uses of argument. Cambridge: 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.

Wise, S. & DeMars, C. (2006). An application of item response time: The effort-moderated IRT model. Journal of Educational Measurement, 43(1), 19-38.

Wodrich, D. L. & Schmitt, A. J. (2006). Patterns of Learning Disorders: Working Systematically from Assessment to Intervention. New York, NY: Guilford.

Wylie, E.C. (2012, June). Using learning progressions to support item and task development for formative purposes in middle school mathematics. Presentation made at the Chief Council of State School Officers (CCSSO) National Conference on Student Assessment, Minneapolis, MN.

Xu, D. L., Yang, J. B., Wang, Y. M. (2006). The ER approach for multi-attribute decision analysis under interval uncertainties. European Journal of Operational Research, 174(3), 1914–43.

Yang, J. B., Xu, D. L. (2002). On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. IEEE Transactions on Systems, Man and Cybernetics Part A: Systems and Humans, 32(3), 289–304.

Ysseldyke, J. E., & Olsen, K. R. (1997). Putting alternate assessments into practice: What to measure and possible sources of data (Synthesis Report No. 28). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis28.htm

Zwick, R. (2007). College admissions in twenty-first-century America: The role of grades, tests, and games of chance. Harvard Educational Review, 77(4), 419-429.