Chapter 5
Simulation Scoring

Eric Sydell, Jared Ferrell, Jacqueline Carpenter, Christopher Frost and Christie Cox Brodbeck

The history of psychological assessment stretches a century past, and until the advent of the Internet, it proceeded at a gradual pace. Now, as connected devices become ubiquitous, the methods we use to collect data are increasingly varied, and the amount of data our field captures is truly vast. For assessment developers, the game has changed—we are less focused on studying the properties of a response scale or particular item type and more concerned with the grand challenge of predicting human behavior. We are at a tipping point, at which our power to collect massive amounts of varied response data will allow us to reach the predictive potential of our field. Simulations are at the forefront of this future.

Although technology-based simulations represent an exciting and engaging future for the testing world, novel item types themselves will not revolutionize our ability to predict important outcomes. With current technology, there is no silver bullet that will significantly improve the size of our criterion-related correlation coefficients. This is not to say that there are no incremental gains that can be made—there certainly are—but we believe major predictive improvements will be made it two areas: (1) combining information across item types and assessment experiences, and (2) leveraging the power of increasingly large sample sizes.

Combinatorial Scoring Although a tremendous amount of research has been directed toward individual scales and item types, vastly less attention has been given

E. Sydell (✉)
Shaker Consulting Group, Cleveland, OH, USA
e-mail: Eric.Sydell@shakercg.com

J. Ferrell
e-mail: Jared.Ferrell@shakercg.com

J. Carpenter
e-mail: Jackie.Carpenter@shakercg.com

C. Frost
e-mail: chris.frost@shakercg.com

C. C. Brodbeck
e-mail: Christie.Brodbeck@shakercg.com

M. Fetzer, K. Tuzinski (eds.), *Simulations for Personnel Selection,*
DOI 10.1007/978-1-4614-7681-8_5,

to how diverse scales and item types interact to predict important outcomes. A high Extraversion score does not indicate that a subject will take every opportunity to speak. The complex human persona demands that if we are to achieve higher levels of predictability, we must take into account the effects of environment, mood, and more. The use of technology-based simulations as stand-alone assessments and the use of simulation exercises in combination with more traditional assessment item types hold promise for increasing predictive power through combinatorial scoring.

Big Data Our field's efforts at combinatorial scoring are drastically curtailed by lack of statistical power. Research is severely limited by small sample sizes. Many statistically significant findings have been reported that fail to hold up in cross-validation samples. However, the good news is that as connectivity and Internet delivery of assessments grow, organizations are increasingly able to provide large sample sizes for both validation projects and ongoing hiring needs. To be sure, many companies still do not collect numerical job performance data to the extent possible, but improvements are constantly underway. The big data movement has barely begun in the human resources arena; but, as it grows, we expect vastly greater ability to determine scoring methodologies that have ever-greater predictive power.

In addition, a fundamental shift is afoot in terms of how candidate data is collected. Whereas for nearly the entire history of our field, data have been collected using a question–response format, simulations are now allowing researchers to directly measure human behavior. In other words, we can now move from asking what a person *would do* in a certain situation, to observing how they *actually behave* in that, albeit virtual, situation. The online worlds being created are constantly becoming more lifelike, and as this occurs, we expect to see continual advances in levels of assessment realism.

How do you score a simulation? This question is unanswerable, as there are as many ways as there are simulations. You must first consider the purpose of the simulation—development, training, selection, etc. Moreover, if for selection, how will the scores be utilized? Will there be cut-off scores, subscores, broad or narrow scores? Once these issues are resolved, you can consider the type of simulation you will develop: Will it be a pure simulation measure or some combination of simulation and traditional item types? Will you seek to minimize adverse impact while also maximizing predictive validity against some type of criterion measure or measures? Will you have access to a validation sample? If so, how large will it be? All of these questions must be answered before determining the ideal scoring methodology.

The scoring of simulations is a dynamic and highly intricate topic. The macrolevel issues above will continue to influence the field for years; but, in this chapter, we also discuss a number of less nebulous topics. These include custom or local scoring models, broad versus narrow scores, automatic scoring of qualitative information, and branching logic.

5.1 Custom Scoring

The typical scoring philosophy among test vendors involves creating scoring routines based on analytical results culled from multiple samples of data. The intent is to guard against capitalizing on local variation by using scoring rules that have been shown to hold up in many different samples. We argue here that if the purpose of your assessment is to describe an individual in universal terms, as it is with many off-the-shelf assessments, this is the proper approach. However, if your purpose is to predict some outcome in a specific environment, then the situation becomes more complex. Too often, assessments that were designed to describe are inappropriately used to predict.

When the goal is to predict job performance outcomes, specificity matters. If the trait of Extraversion has a metaanalyzed validity coefficient of 0.3, does that mean it will predict for the role you are studying? One might surmise that it would be predictive of bank tellers' sales success, and yet at Shaker Consulting Group, our consultants have found a much less intuitive predictor to be a vastly greater and more stable indicator of sales success: computer skills.

If we compare two different sales positions, one a door-to-door sales position and the other an inbound call center position, we do not expect the trait of Extraversion to predict results identically. In the outside position, the primary relevant personality trait indicative of success may well be Extraversion, for it is incumbent upon the individual to assert him- or herself to unsuspecting and likely closed-minded potential customers. In the inbound role, Extraversion scores will likely not predict success at all, as the individual simply needs to answer the phone and ask scripted questions about whether a caller would like to purchase their products or services.

At Shaker Consulting Group, we create custom simulation modules for particular roles, but we typically hold constant the core measurement features of the simulation across clients. For example, a multitasking exercise might be customized to reflect the specifics of a role or organization with respect to the type of widget involved in some stacking exercise or the type of information shown in a call queue display. However, the core measure involving a numeric calculation coupled with a simultaneous task, such as clicking a button to take a new mock call, can well remain the same across different versions of the exercise. The validity evidence of the core measure is assessed and combined from many instances of administration, attesting to the stability of the measure.

One of the most valuable features of technology-facilitated simulations is the ability to include and deliver a wide variety of job-related assessment activities to candidates in a manner that more realistically depicts actual job tasks. The use of custom scoring with simulation-based assessment activities allows us to meet the goal of creating a high fidelity, realistic experience for candidates while also capitalizing on the rich data obtained from simulation exercises to predict on-the-job performance more precisely. Although the value of variety and realism of technology-facilitated simulations has led to an increased focus on creating novel, realistic exercises to assess different facets of performance, an important issue that we cannot afford to

overlook involves decisions about the ways to summarize data to present candidate scores.

Researchers have recently examined whether meta-analysis, local study, or Bayesian analysis is the most accurate way of estimating local validity (Newman et al. 2007). Contrary to some conventional wisdom, local studies can actually provide more accurate estimates of predictor validity in some circumstances than meta-analytic methods. However, the best way to estimate the validity of a custom scoring procedure may be to combine the local estimate with Bayesian prior probabilities generated from meta-analytic results. Using this technique, selection scientists can weight a local validity result with the prior meta-analytic evidence to arrive at a more stable validity estimate (or Bayesian posterior).

In the scoring arena, one of the most common debates deals with the relative effectiveness of generating and reporting broad versus narrow competency scores. In order to assist test developers in deciding whether broad or narrow competency scores are most appropriate for a simulation-based assessment, the benefits and drawbacks of each as well as the best practices for implementation are discussed in Sect. 5.2.

5.2 Competency Scores: Broad Versus Narrow

5.2.1 Broad Performance Competency Scores

At a high level, simulations can be developed to yield specific scores around constructs such as multitasking ability, typing speed, cash transaction accuracy, and many other narrow scores. These may well be predictive of certain aspects of on-the-job performance. However, the present authors argue that simulations can be leveraged much more broadly to predict not only specific task performance factors, but also higher level competencies that represent a sizable portion of the job performance domain. In fact, more than other assessments, simulations offer the potential to predict overall job performance due to their ability to include diverse item types and provide a more realistic experience. Simulations offer the ability to measure a candidate in a more holistic fashion than an assessment developed around a specific item type or construct.

Broad competency scores combine multiple narrower facets and dimensions from either the same or different simulation exercises into more general, all-encompassing output reports. The goal of this combination effort is to utilize customized combinations of exercises to explain performance better than is possible through the utilization of narrower facets. These broad composites, highly popular in the world of customized assessments and simulations, are often created to align with an organization's own competency language, making them more easily interpretable by lay clients not trained in the technical aspects of personnel selection. For example, a broad competency score could be computed to predict overall performance based on a simulation, instead of reporting a score from each specific aspect of the simulation.

5.2.1.1 Benefits

A significant benefit of creating broader competency scores is that this practice allows key variables to be at the center of attention (Smith 2002). For example, a broad scoring composite could be created to measure expected overall job performance based on performance on a simulation. This has the benefit of putting performance at the forefront of attention when decision makers within the organization look at output reports.

Broad performance competencies also afford decision makers with increased efficiency not offered by narrower competency scores. Looking back at the example given above of a broad performance composite, this allows for an easy rank ordering of candidates on a key variable or variables of interest. As described by Hattrup (2012), no matter how many exercises are on a selection assessment, ultimately the decision boils down to a dichotomous choice between hiring and not hiring the candidate. Thus, this broad scoring methodology makes it much easier for decision makers to see the big picture instead of being caught up in an overabundance of narrow facet scores when drawing conclusions from assessments.

Multiple studies have shown that the use of composites can help to decrease the potential for an assessment to exhibit adverse impact (e.g., Bobko et al. 2007; Sackett and Ellingson 1997; Schmitt et al. 1997). This is generally due to the compensatory nature of broad competencies, in which case different scales all combine in a certain way to provide valid assessments with minimal risk for adverse impact. Other researchers have taken a more technical approach to finding optimal weighting schemes for maintaining high levels of validity while minimizing the risk for adverse impact.

De Corte and colleagues, in a series of studies (i.e. De Corte 1999; De Corte et al. 2007), took a technical approach to find the pareto-optimal tradeoff between validity and adverse impact. The pareto-optimal tradeoff attempts to solve the diversity–validity dilemma, wherein assessment providers are constantly in a tug of war between trying to provide the highest levels of validity while also minimizing the potential for an assessment to result in adverse impact—two tasks at odds with each other (De Corte 1999; Pyburn et al. 2008). The term "pareto-optimal," born from economics literature, refers to situations in which the increase of one factor is at odds with another factor. Applied to selection, increasing validity is at odds with decreasing the risk of adverse impact, and thus trying to optimize one inherently works at the expense of the other.

De Corte (1999) initially proposed a constrained nonlinear methodology for creating composites to minimize adverse impact concerns while working to maximize performance gains from a selection system. In De Corte's methodology, a constraint was placed into the weighting equation, setting the minimum acceptable adverse impact ratio. This value became a key consideration in the formation of the weights. A limitation of De Corte's methodology was that instead of optimizing both variables, the equation only optimized one (validity) while constraining the other (adverse impact potential). This issue was addressed by De Corte et al. (2007), in their article, which provided a new formula for the calculation of pareto-optimal tradeoffs between adverse impact and validity. The updated model presents numerous points

that show differing levels of validity and adverse impact potential, based on differential weighting of the scales within the broad competency. One point included is the pareto-optimal level for validity and adverse impact in conjunction with each other.

The weighting equations described above provide evidence of examples where differential weighting of different subfacets within a broad composite allows for a compensatory system in which validity and adverse impact are optimally balanced. Scales can be weighted together in formations such that a single scale or exercise that is not only highly valid, but also at a higher risk for adverse impact (e.g., a cognitively loaded measure) can be combined with other predictors (e.g., personality) that are at a lower risk of violating adverse impact ratio cutoffs, creating a composite that is highly predictive of performance and also adheres to federal regulations on adverse impact.

5.2.1.2 Drawbacks

While there are numerous advantages to the utilization of broad scoring composites, there are also some distinct drawbacks to this practice, depending on the situation. In certain instances, broad competency scores may actually serve to disguise serious failings of candidates due to the inherent compensatory nature of the scoring system. For example, a candidate could score well on a broad composite without raising red flags that would be more likely to present themselves with narrow competency scores. Along the same lines, broad composites can potentially disguise the factors in a simulation that are the key drivers of performance making it hard to home in on key ways to increase organizational effectiveness.

Weighting issues, while possessing the potential to benefit a system in certain situations, also present potential drawbacks to the utilization of broad competencies in simulation scoring. As composites begin to integrate more factors, weighting of individual scales or exercises becomes a critical issue. The problem lies in the fact that weighting is not straightforward, thus subjectivity becomes injected in the scoring systems (Gatewood et al. 2010). For example, there are four main schools of thought on the weighting of predictors to form composites: (1) regression weighting, (2) reliability weighting, (3) a priori weighting, and (4) unit weighting (Hattrup 2012). The issue lies in the fact that there is much disagreement in the literature regarding which weighting scheme is optimal, leaving practitioners in a precarious position in terms of having to make and support a weighting decision. Doverspike et al. (1996) advise practitioners, no matter which scheme they utilize, to describe in detail the process implemented in the decision making regarding weights within a composite.

The final major drawback to broad performance composites strongly parallels any overly large organizational intervention. The issue is that when many parts are intertwined, a major change to one of the facets within a competency score could cause a host of weighting and predictability issues in the composite as a whole, leading to an entirely new set of headaches for practitioners. As such, practitioners need to be cognizant of this when choosing to utilize broad competency scores and ensure they understand the effects changing one part of the simulation can have on the properties of the competency score(s) as a whole.

5.2.2 Narrow Competency Scoring

Narrow competency scores are generally comprised of separate facets for different exercises and competencies and are presented as different metrics on the simulation output reports. For example, problem-solving skills would be separated from personality facets in the output report instead of potentially being combined to create an overall performance score as they might be in a system implementing broad competency scores. This approach to scoring has its own benefits and drawbacks, discussed in Sect. 5.2.2.1 and 5.2.2.2.

5.2.2.1 Benefits

The primary benefit of using narrow scales is their ability to present results for each facet or section in a more straightforward manner than their broad competency counterparts do. The results are much more transparent than many of the broader composite scores. This has numerous positive implications for practitioners, including ease of theoretically linking predictors with narrow criteria dimensions, increasing the ease of showing rationale behind inclusion of assessment aspects (Arthur et al. 2003; Christian et al. 2010).

The narrow scoring of facets also makes weighting less of an issue in most cases. This is because narrow competency scores generally do not require differential weighting of different simulation activities, instead, commonly requiring unit weighting of items into each narrow composite. This reduces the potential for subjectivity in the initial weighting of the competencies and in certain cases may make a system less vulnerable to legal action based on the scoring methodology.

Narrow composites offer the potential for more direct feedback than their broad counterparts do. The narrow composite approach is well suited to discovering and illuminating what specific facets actually drive performance while offering feedback that focuses on those specific facets. Moreover, they can be equally beneficial for the reverse situation, in which it may be important for decision makers to flag candidates for serious deficiencies on specific competencies or simulation exercises that have been shown to be critical to organizational success. The effects of this, on deciding which type of scoring competency to utilize, will be discussed in the best practices section (Sect. 5.2.4).

5.2.2.2 Drawbacks

While there are benefits to having more detailed output reports that include multiple facets, this method is not without its own drawbacks. The drawbacks here are often due to the interpretation of numerous narrow composites. Although broad composite scores are computed via an actuarial manner, wherein there are hard numbers to back up decisions, if care is not taken with training end-users on the meaning of various narrow facets, it is easy for hiring decisions to be based on softer interpretations,

which may be more difficult to defend after the fact (Grove and Meehl 1996). This ties back to the tradeoff in which broad composites throw subjectivity into the creation of the composites themselves and narrow composites inject subjectivity into the potential for differential interpretations of the same set of scores, thus necessitating some sort of output report training to try and mitigate the potential for this to decrease the utility of the simulation as a whole. Moreover, the narrow facets, while interpretable to developers of assessment content, may not be interpretable to lay end-users. This issue ties into the concern about inconsistent interpretation of output reports across key decision makers. As an example, decision makers may not completely understand what a narrow facet, such as Extraversion, directly means, or how it would specifically relate to performance, causing different interpretations depending on who is reading the output report of a candidate.

Another concern with this approach is the potential to accumulate too many narrow facets or composites on a scoring output, leading to information overload, and thereby decreasing the administrative efficiency of making decisions from the simulations. Indeed, it is very common for simulations to have upward of 30 or more different narrow scales or composites, which can quickly become a nightmare for decision makers within an organization. As such, practitioners should be careful to focus on key composites that drive performance; in addition, they should even potentially consider removing more peripheral scales or composites that focus on predicting extremely narrow subsets of performance.

5.2.3 Psychometric Considerations

It is a fundamental element of psychometrics that you cannot have validity without reliability. This general dictum has been ingrained into graduate students' brains for ages (along with the idea that correlation does not imply causation). However, there is vast misunderstanding of the nature of the relationship between reliability and validity.

Theoretically, a measure must be reliable in order for it to be valid; however, in practice, it is extremely difficult to verify this relationship. The vast majority of scale development utilizes coefficient alpha as the reliability estimate of choice due to its simple computation. However, internal consistency is but one type of reliability estimate, and while internal consistency is important for scale interpretability, what if the scale is combined with other items or scales to yield a broader competency score? A heterogeneous scale may still be a reliable indicator of relevant characteristics. When the purpose of a simulation is shifted from description to prediction of real-world outcomes, interpretability is less important than high predictive power.

Many continue to focus on coefficient alpha for its ease of use. However, when creating broad competencies, we recommend following the newer approach of Linear Composite Reliability (Nunnally and Bernstein 1994). Although not discussed here, this approach provides an estimate of reliability that takes into account the

reliabilities, relative weight, and variance of each component within the composite as well as the overall variance of the composite.

5.2.4 Implementation of Broad or Narrow Competency Scores: Best Practices

It is prudent to disentangle the benefits and drawbacks of broad versus narrow competency scores through explaining situations better suited to one methodology over the other. This section will begin with situations better suited to broad scoring composites, followed by an examination of situations in which a narrow scoring composite is the more appropriate choice.

5.2.4.1 Situations Best Suited to Broad Versus Narrow Scoring Composites

There are numerous situations in which either broad competency scores or narrow competency scores are better suited to achieving the goals of the simulation through which they are derived. Factors influencing the relative effectiveness of broad versus narrow competency scores are numerous, and thus an exhaustive list of situations is beyond the scope of this chapter. Nevertheless, there are certain general cues that can help practitioners decide whether to generate broad versus narrow competency scores based on a candidate's performance on a simulation. These clues can come via situational constraints, the purpose of the simulation itself, and the nature of the outcomes the simulation is designed to predict.

As discussed previously, situational constraints can determine the optimal composite construction methodology. One such organizational constraint deals with the time allotted to make decisions based on assessments relative to the number of people who complete an assessment. Based on a sheer lack of available time by key decision makers, the efficiency in decision making is often vital in organizational settings. For example, imagine a company that administers a simulation to thousands of candidates for a small number of job openings. It would be overwhelming for hiring managers to sift through report after report littered with narrow competency scores. Instead, a broad performance composite score would be ideal here, as the hiring managers could utilize applicant-tracking databases to sort candidates based on how well they are predicted to perform overall in this specific work environment. This would make deciding which candidates to advance to the next stage of the selection/promotion process much easier and more straightforward as opposed to trying to compare thousands of candidates on numerous narrow facets.

The purpose of the simulation can also be used as a deciding factor in whether to create broad or narrow competency scores. Although each situation will be different, there are some general situational factors that can affect whether broad or narrow competencies would be better suited. As discussed above, broad competency scores can increase the efficiency and uniformity of decisions, and therefore can often

times be more practical for decisions being made in a hiring/promotion context, especially one with a high volume of candidates. On the contrary, the specificity of narrow competency scores allows for the ability to understand and alter specific behaviors and is well suited for developmental exercises in which it is valuable to be able to pinpoint specific opportunities for future skill enhancement, whereas broad competency scores would only be able to identify if there is a gap in performance at a much more general level, thus not being able to give specific suggestions for improvement.

The nature of the criteria may also affect the decision of whether it would be optimal to utilize broad or narrow scoring composites. Ideally, the goal is to match the criteria with the predictors, such that if the criterion is broad, a broad composite would be viewed as optimal, and if the criterion is narrower in nature, the scoring composite should be narrow to match as well. As an example, if the criterion of interest is organizational performance, it would be more advantageous to have a broad performance composite than to generate a plethora of narrow composites and expect decision makers to wade through the information and draw conclusions. Conversely, if the criterion of interest is communication skills, a narrow composite composed of exercises that tap this factor is going to be more beneficial than would an overarching performance composite.

While certain examples have been given in which broad or narrow competencies are preferred, it is also often the case that both are utilized in congruence with each other. For example, numerous simulations generate output reports that include broad competency scores and more narrow scores to help reap the benefits associated with the utilization of both. This methodology is beneficial because it allows decision makers to be able to employ a cursory screening of unqualified candidates as well as a more in-depth comparison of qualified candidates before proceeding to the next step in an organizational decision-making process.

Up to this point, we have discussed scoring considerations relevant to any assessment employing simulation exercises. As discussed in this chapter and others in this book, the array of simulation exercises used in assessment is vast and varied, thus it is beyond the scope of this chapter to cover scoring considerations specific to each type of simulation exercise used in assessments. However, we do want to highlight scoring considerations pertaining to some particular innovations in simulation exercises and test construction. The following section will discuss the use of automatic scoring in computer-based simulations and the techniques associated with the application of this methodology to simulation construction and scoring.

5.3 Automated Scoring of Qualitative Data from Simulation Exercises

An exciting innovation offered by technology-facilitated simulations is the opportunity to collect and automatically score open-ended responses from candidates. Compared with cumbersome, essay-style assessments of the past, simulations allow

organizations to collect writing samples from candidates in novel ways and tailor the stimuli to job-specific situations. For example, a candidate may be presented with a hypothetical situation or problem and be required to generate multiple possible solutions or strategies to solve this problem. Embedding open-ended items within simulation exercises allows organizations to gain a more comprehensive understanding of the candidate's thought processes and complex problem-solving skills relative to a simple Likert-type scale item (Ackerman and Smith 1988; Birenbaum and Tatsuoka 1987). Furthermore, the range of possible responses is virtually unlimited, providing additional information that may be particularly useful when attempting to select candidates for higher level positions such as managerial or leadership positions (Zaccaro et al. 2000).

Until recent technological advances, the benefits of including qualitative elements in selection assessments were overshadowed by administrative impracticalities. Prior to the advent of automatic scoring methods, the time demands required to evaluate these responses were extensive, not to mention costly. Consider that each response must first be generated by the candidate before an appropriate scoring system is developed; responses must then be read in their entirety and finally scored. In addition to the time requirements, this scoring methodology is vulnerable to rater errors (Zaccaro et al. 2000).

The development of an automated scoring process promises to mitigate (if not eliminate) the disadvantages associated with qualitative scoring, while also maintaining the measurement benefits of this method. Reducing the time requirements in what is typically an extremely time-intensive process is the most obvious advantage of applying an automatic scoring methodology to qualitative data. Without an automated scoring process, multiple reviewers or raters would be required to read and score each piece of writing. An automated process eliminates this time consuming endeavor as all pieces of writing can be scored instantaneously. In addition to the reduced time requirements, automatic scoring processes introduce an increased level of objectivity to the scoring of qualitative data. A third advantage of applying an automated scoring process to written text is the comprehensive nature of the evaluation. Regardless of how efficient a human rater is or how closely they read a written sample, they will not be able to remember every word that they read and factor it into their final evaluation. Automated scoring methods by contrast, are capable of evaluating each piece of text in the response and using each word to develop a refined scoring process.

5.3.1 Automatic Scoring Methods

To date, automatic scoring methods have been applied more frequently in the education context (i.e., evaluating student essays, ACT, GRE) than within the selection domain (Attali 2004; Burstein and Chodorow 1999). Results from research in the education domain generally indicate that these automatic scoring methods can reliably

reproduce human ratings, and in some cases, the automatic scoring methods actually appear to be more accurate than a human grader (Burstein and Chodorow 1999; Shermis 2012). Although there are numerous automatic scoring software programs available, the programs tend to be conceptually similar to each other[1] (Fielding and Lee 1998; Tesch 1990). However, these automatic scoring software programs vary somewhat in their methodology. Some programs utilize essays that were previously scored by human raters. These essays are divided into groups (e.g., high scores, average scores, and low scores), and the program is then trained to recognize the key differences between the essays in each group. In addition, many of these software programs evaluate writing samples based on grammatical properties such as subject-verb agreement, sentence completion, and punctuation.

Evaluating grammatical properties is not only popular in the educational domain, but it is also utilized in other contexts, including employee selection. The primary advantage of utilizing these types of techniques is that they are generalizable to nearly any context. However, there is perhaps a great deal more that can be uncovered by examining factors other than grammatical quality.

5.3.2 Measuring More Than Essay Quality

Utilizing automatic scoring to examine constructs beyond simple essay quality may be particularly valuable for organizations when assessing candidates. Consider that for many jobs composing a grammatically sound sample of writing may not be something an employee is required to perform. However, with technology-facilitated simulation exercises, qualitative item types can be used in novel forms to create a higher fidelity experience that taps numerous job-relevant constructs. There are some encouraging research results relating open-ended items to personality traits, leadership characteristics, and coping styles. For example, research on automatic scoring systems has shown their ability to predict personality traits, leadership characteristics, and individual coping styles based on qualitative response characteristics such as word choice (Fast and Funder 2008), simple word count (Hirsh and Peterson 2009; Lee and Cohn 2009), and idea complexity (Dudley and Cortina 2008). Assessing numerous constructs allows for job specific customizable scoring that achieves maximum validity. These benefits do not come easy, as there are numerous challenges associated with developing an automatic scoring system for qualitative items.

In general, when attempting to measure constructs beyond basic essay quality, scoring development is somewhat complex. For example, identifying words and phrases that indicate a high standing on a particular construct is not a simple process. Consider an assessment intended to measure a trait such as Conscientiousness. To develop the automatic scoring system, it first must be determined which words and phrases are more likely to be used by a person high on conscientiousness compared

[1] It is estimated that eight automatic scoring methods (AutoScore, LightSIDE, Bookeete, E-rater, Lexile, Project Essay Grade, Intelligent Essay Assessor, Crase, and IntelliMetric) represent approximately 97 % of all of the automatic methods used today to evaluate student essays (Shermis & Hamner 2012).

with a person low on conscientiousness. Although this task in itself is complex, it is made more so when considered in a selection context. As previously discussed, free form writing samples have generally been utilized to demonstrate the effectiveness of automated scoring methods to predict personality. When a candidate is composing a written sample for a potential employer to examine, they are not using a free form writing style. Candidates will generally try to put their best foot forward during the assessment process and therefore the range of responses is extremely restricted when compared with free form written samples. Therefore, the amount of words that will effectively differentiate among candidates on a particular construct is likely to be reduced in assessment contexts.

5.3.3 Automated Scoring of Qualitative Data: Future Directions and Suggestions for Practice

While identifying certain keywords and phrases to train an automated scoring system is not an easy task, there are software programs that can help to facilitate this process. One user-friendly program that can be utilized during this process is the Bayesian Essay Test Scoring System (BETSY; Rudner and Liang 2002). To use BETSY, one would divide writing samples into groups based on a particular criterion. For example, writing samples from highly extraverted individuals would be placed into one group while those from low extroverted individuals would be placed into another. The BETSY software would help to identify words that are used more frequently by members of the high Extraversion group. Users would be asked to identify words that differentiate between groups and are theoretically appealing. Eventually, these words and phrases can then be uploaded into software that identifies keywords, or BETSY could actually be used to score these written samples.

The BETSY software, along with many other software programs, utilizes a simple word count method. Initially, this word count function may appear overly simplistic. However, with respect to personality traits, there are numerous reasons why one would expect simple word choice to be related to candidate personality (Fast and Funder 2008). Furthermore, when utilizing a simulation that requires candidates to recall specific information from a particular passage, word count procedures can be very useful in scoring (Mumford et al. 2000).

There are some important issues that must be considered before implementing a word count procedure as part of the scoring for a simulation exercise. For example, consider an assessment that is designed to measure Achievement Orientation. It may be that words such as “motivated” and “driven” are identified as more likely to be used by a person that is high in Achievement Orientation. However, if a person uses a word such as “inspired” rather than “motivated” they would not receive an increased Achievement Orientation score despite the fact that they were attempting to convey the same message as another person who may have used the word “motivated.” Thus, careful consideration must be given to issues such as whether to score certain synonyms theoretically linked with the words identified by the word count software.

Automated scoring of qualitative item types is merely one example of innovative item types used in technology-facilitated simulations. Technological advances have

also allowed for innovation in the construction of simulation-based assessments. For example, test developers now have the option to choose whether to use a nonlinear testing approach in place of a traditional linear testing approach, in order to provide a customized experience for candidates and achieve more precise measurement. Instead of presenting the same content to all candidates, nonlinear approaches such as branching logic or Computer Adaptive Testing (CAT) provide candidates with a customized experience based on their individual test performance. Applying nonlinear techniques to simulations for employee selection can enhance many of the benefits of simulation-based assessment.

5.4 Branching and Adaptive Testing

While still a relatively young approach to testing, nonlinear testing approaches are gaining popularity. The military introduced the first large scale CAT, the Armed Service Vocational Aptitude Battery (ASVAB) in 1982. CAT is a common type of nonlinear test design in which the test adapts to the candidate by successively presenting items representing a higher or lower level of the test construct based on the candidate's previous response (Drasgow and Olson-Buchanan 1999; Wainer 2000). For example, if a candidate incorrectly answers a mathematical ability item, he will be presented with an easier mathematical problem next. After multiple iterations of presenting items based on performance on the previous item, the test algorithm is able to pinpoint the candidate's level of the construct being measured. Since the introduction of the ASVAB, CAT has been increasingly used for professional and licensure examinations as well as academic entrance examinations such as the Graduate Record Examination (GRE). More recently, private sector companies have begun introducing nonlinear assessments to their employee selection processes. Beyond CAT, which may be primarily used when the goal is to shorten a test while maintaining or enhancing measurement precision, other nonlinear test approaches may be used to create branching or storyline experiences for candidates. The goal is to adapt the test to the candidate's performance in order to create a realistic simulation that enables more sophisticated modeling of potential performance. Both methods can be applied to simulation scoring and thus are discussed in the following sections.

5.4.1 Benefits of Nonlinear Testing

Nonlinear testing approaches can offer several enhancements to simulations for employee selection. Although typically associated with multiple-choice knowledge-based testing, nonlinear approaches can be leveraged across a variety of constructs and measurement formats. Techniques such as work samples, problem solving, and situational judgment tests can be implemented using various response formats. Beyond multiple-choice, candidates can interact with the content in ways that allow for more refined data collection such as clicking on a portion of a graph, dragging

and dropping content, using a sliding bar, or manipulating content to build and create (Bejar 1991; Clauser et al. 1997). In addition, multimedia approaches—audio, video, animation, and a variety of imagery—can be easily incorporated. For example, a measure of mechanical ability can require candidates to view a video or image and then click on the part of a machine most likely to be causing a problem. Furthermore, because nonlinear testing allows for the customization of content based on previous responses, a candidate's problem-solving process can be explored in a way that mimics on-the-job scenarios. After selecting the machine part, candidates can be prompted to select which tools they would use to investigate the problem or to choose their next step, if the first hypothesis is incorrect. Storyline branching approaches such as these create an interactive and unique experience for candidates while also serving as a preview of the job. Importantly, they also give richer and more job-relevant data to potential employers. An added benefit for employers is that, unlike traditional approaches to evaluate branching or adaptive testing such as assessment centers, this complex information can be scored instantaneously.

The growing popularity of nonlinear testing can be attributed to several psychometric benefits touted over traditional linear tests, particularly CAT. By presenting targeted items based on a candidate's performance, fewer items are needed to zero in on the candidate's level of the measured construct. Candidates spend less time on items that are too easy or difficult, and more time on items that enable the fine tuning of their scores (McBride and Martin 1983). By reducing test length, test security may also be enhanced. Each candidate receives a different version of the test and is exposed to fewer items, making it less likely that test content will be shared or compromised. Even when compared with alternate forms of traditional tests, CATs help prevent cheating by giving a slightly different version to each candidate (Guo et al. 2009). This is especially beneficial for unproctored testing where content can be more easily compromised.

CATs have also been shown to enhance the reliability and validity of a test. Shortening a test through CAT also reduces fatigue associated with long tests (Tonidandel et al. 2002). Without fatigue effects clouding construct measurement, CAT can produce assessments that are more reliable and valid than linear tests despite their shorter lengths. Gains in validity also stem from the greater precision of adaptive tests. Because test content is successively targeted to candidates' individual performances, more information can be gathered about their specific ability levels. Although linear tests are often best for measuring medium levels of ability, CAT's customized content means that the test can quickly target test content for a wide range of ability levels (McBride and Martin 1983).

Importantly, adaptive testing is typically perceived positively by candidates as well. Using adaptive testing techniques to create simulations and storyline experience can feel more realistic and face valid for candidates (Hanson et al. 1999; Parshall et al. 2010). Proctor & Gamble also found in an extensive implementation of CAT that candidate perceptions of the appropriateness and fairness of the CAT were comparable to traditional tests (Gibby et al. 2008). In addition, the shorter length reduces test fatigue and the time needed to complete the test, making the test less taxing on the candidate. Research on the effects on test anxiety has been less

consistent. Although anxiety is often reduced for those of medium and low abilities, adaptive testing may increase anxiety for those who do well early in the test as they will be quickly presented with difficult items (Tonidandel et al. 2002).

5.4.2 Complexities and Considerations

Although there are many benefits of nonlinear testing, this approach should be carefully considered before implementing. One important consideration is the technology comfort level of the intended test audience. Computer based testing in general may not be appropriate for audiences unfamiliar with the technology, and the expected computer proficiency of the target group should be considered when designing the interface of the test (Parshall et al. 2010). Creating an overly complex or customized test experience may limit the usability of a test.

In addition, the complexities and resource requirements inherent in this approach mean it is not always appropriate and must be carefully implemented. Implementing and maintaining an adaptive test is time and resource intensive. In the case of CAT, creating the test requires a large pool of items representing all levels of the targeted construct and a large sample to validate those items. Depending on the target group for the test implementation, this approach may or may not be suitable. For example, for small target groups, the expense and resources to develop the test may not be justified.

Complexities also arise in item development and the creation and implementation of complex scoring and item presentation algorithms. Mistakes in the development process can greatly impact test validity. Items must appropriately represent all levels of the construct, and algorithms must accurately choose and score responses based on previous responses. For example, a balance must be maintained between creating a short and statistically sound test and fully representing the construct domain. Too much attention to time and statistical considerations can result in narrow measurement of the construct domain for some candidates and create nonparallel forms of a test that do not allow candidates to be accurately compared (Huff and Sireci 2001). Implementing branching logic or adaptive techniques in simulation-based selection assessments requires rigorous item bank and test structure development as well as consideration of several scoring options.

5.4.3 Assessment Design and Scoring Considerations

The psychometric foundation guiding item calibration, item selection, and scoring for CAT is Item Response Theory (IRT). IRT advanced by Rasch (1960), Birnbaum (1968), and Lord (1970) describes the relationship between observed test performance and a test-taker's underlying ability on a particular trait. Typically, one or more item parameters (e.g., item difficulty, discrimination) are combined within a

logistic function (Folk and Smith 2002; Lord 1980; Wainer et al. 2007). Although the simplest of models, the one-parameter logistic (1-PL) model considers just one item parameter, a three-parameter logistic (3-PL) IRT model is common among modern CAT designs.

The chosen model informs the calibration of test items, such that fitting an item response model to pretest data allows for the recording of each item's estimated parameters (Wainer and Kiely 1987). IRT also provides the basis for adaptively selecting test items based on an examinee's response to previous items (the item selection algorithm), and scoring the adaptive test as a whole.

Due to the heavy demands placed on test items in nonlinear testing environments, one of the costs of developing adaptive tests of any sort is the writing and calibration of items. Item pools for single-construct, IRT-based adaptive tests are necessarily large in order to cover the range of ability assessed and provide a sufficient amount of alternative items for the purposes of test security. In developing items for adaptive test formats, tests developers must consider the range of ability in the candidate pool, content coverage, and adhere to the assumptions of item parameter estimation models (Flaugher 1990; Hambleton 2002). Large pools of items also necessitate a substantial pool of trial items and large samples of pilot test-takers (Zickar et al. 1999).

Methods for adaptive item selection must be chosen with consideration of several test features, including psychometric characteristics, content specifications, and item exposure (Folk and Smith 2002). Popular item selection algorithms used in modern adaptive test designs are often based on the psychometric selection criteria of maximizing information about the examinee's current ability level (Folk and Smith 2002; Lord 1977; Rasch 1960). Alternatively, Bayesian item selection methods, such as the one described by Owen (1975), are also often used. Beyond these, other models for automated item selection have been developed with the goal of improving the balance between psychometric efficiency and content requirements (e.g., the weighted deviations model, Stocking and Swanson 1993; optimal constrained adaptive testing, van der Linden 1998). Van der Linden and Pashley (2010) provide detailed technical overviews of several modern alternatives to traditional adaptive item selection models. Item selection models often also include provisions for systematically monitoring item exposure for the purposes of item pool longevity and test security. For more details regarding item exposure, test security, and item pool maintenance, see Davey and Nering (2002) as well as Segall and Moreno (1999).

Test developers must consider multiple options for estimating final scores in order to achieve the optimal balance between precision, simplicity, and fairness. Administrators often confront difficulty in achieving a simplistic scoring model that can be explained to candidates with complex nonlinear tests. Scoring for adaptive tests depends upon test delivery methods and scoring is affected by how examinee responses are modeled. Therefore, developers must be particularly attentive to the manner in which the item parameter estimation and item selection algorithms impact final scores. For example, Owen's Bayesian ability estimate, tested for use in the CAT-ASVAB to update provisional scores after each item and provide a final score, has the undesirable property of providing final scores that depend on the order

in which items were administered (Segall and Moreno 1999). That is, two candidates who answered the same items with the same responses may end up with different final Owen's ability estimates if they received the items in a different sequence.

Adaptive test design features, such as test length and response time modeling, are also important considerations relating to the generation and use of examinee scores. Adaptive tests can take the form of fixed or variable length. For variable length tests, two stopping rules may be used. A target standard error of measurement can be determined and additional questions are presented to the candidate until the target is met, or the tests can be stopped when a specified level of confidence in the pass/fail decision is met (Bergstrom and Lunz 1999; Kingsbury and Weiss 1983; Segall and Moreno 1999). Test developers must consider the incremental informative value of each additional item to determine whether fixed or variable length is appropriate for a particular assessment (Segall and Moreno 1999).

During the creation of the CAT-ASVAB, developers discovered a trend in candidate total time that was opposite of the anticipated response times from the pencil and paper version of the tests. That is, higher ability candidates were spending more time because they received more difficult questions requiring more time to answer (Segall and Moreno 1999). Related to the imposition of time limits, test developers and administrators must consider whether penalties should be imposed for incomplete tests. The necessity of such a penalty will depend on the particular scoring procedure applied to an assessment. For example, the Bayesian scoring procedure used in the CAT-ASVAB contained a bias such that a low-ability candidate could increase his or her score by answering the minimum number of items allowed, taking advantage of estimates that are too close to the population mean. Through a series of assessment simulations, the developers of the CAT-ASVAB settled on a penalty procedure that produces a final score that is "equivalent (in expectation) to the score obtained by guessing at random on the unfinished items" (Segall and Moreno 1999, p. 48). For more details on the penalty procedure implemented in this example, and time limit considerations for adaptive tests, see Segall and Moreno (1999) as well as Schnipke and Schrams (2002).

Building and adhering to the assumptions of item selection and scoring algorithms for adaptive tests can be a complex endeavor requiring extensive resources. Thus, it is important that the items and test format are designed and administered in accordance with measurement goals and the ultimate test purpose (Luecht and Clauser 2002). For simulation-based selection assessments, test developers must consider whether single item IRT delivery and scoring methods adequately fit the nature of the measurement experiences and responses captured through complex computer simulation examinations. As discussed in other chapters in this book, complex interactive exercises, such as those found in computerized simulation assessments, are a potential source for vast amounts of data. Luecht and Clauser (2002) discuss scoring methods for various complex computerized tasks (e.g., correcting embedded errors in an essay passage, producing mathematical expressions that represent a stimulus relationship, and managing patient information through data entry on order sheets) and the challenges of modeling the raw data for such complex tasks. Evident from

this discussion and others (e.g., Wainer and Kieley 1987; Wainer et al. 2006) is that simulations containing complex computer-based tasks may be best suited for adaptive testing models that allow for modeling data representing multidimensional abilities and/or account for associations among subtasks.

Wainer and Kieley (1987) first introduced the concept of "testlets" as "a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow" (p. 190). Wainer et al. (2007) proposed the use of testlets as the unit of construction and analysis in computer adaptive tests to alleviate such difficulties as context effects, item ordering, and content balancing which exist in most traditional algorithmic methods of test construction. Broadly, Wainer and Kieley (1987) propose that testlets can help in two ways: first, by allowing the test developer to recover more control over the test structure that is relinquished with automatic test construction algorithms; and second, by increasing fairness, such that scores for candidates of similar proficiency will be derived from tests of very similar content.

Testlet-based designs allow for the measurement of knowledge in several different content areas (Wainer and Kiely 1987; Wainer et al. 2007); they also allow interdependencies among sequential items referencing single stimuli (occurring, for example, in such situations that involve a large stimulus with several follow-up items; Wainer et al. 2007). Testlet-based designs are useful for allowing for sets of items with multiple response types within one test (Wainer et al. 2006), as is typical of many simulation-based selection assessments.

Developers of adaptive simulation-based assessments should also explore modeling options designed with the intent of allowing for measurement of multidimensional attributes and models that allow for generation of profiles of multiple components related to subtasks and sub processes. Several examples of such complex adaptive tests are provided in Williamson et al. (2006) as well as Segall (2010). In addition, Mulder and van der Linden (2010) discuss test-modeling options for multidimensional adaptive tests in detail.

5.4.4 Potential Future of Scoring Methodologies

Another area of interest to practitioners and academics alike deals with the vast potential for advanced and adaptive scoring methodologies for prehire simulations. Instead of simply combining and weighting the responses given in an assessment to yield a final score, more advanced methodologies may utilize detailed theoretical models that attempt to explain how a particular response to a simulation stimuli relates to other responses in the assessment as well as in other criteria.

Although much research has been conducted on this topic in other fields such as training evaluation, this is a vastly unexplored area to this point in scoring simulations for selection purposes. The ever-improving realism and complexity of prehire simulations, however, will likely push this research area to the forefront of the assessment field in the very near future. Indeed, one of the most exciting promises held

by ever-improving simulations is the move from self-report, generic assessments to more performance-based, autonomous, and customized experiences for participants. As this move occurs, it is vital to be able to still gather information about higher level, complex, and even sometimes abstract abilities of candidates based upon observations of the lower level, concrete behaviors they perform within these high fidelity environments. Everything from the length of time a candidate spends doing something in a simulation to the amount of information they gather before acting in some manner within the simulation to their exploratory patterns can be of value in telling us more about a candidate, yet without improved scoring systems, much of this information may be wasted due to inadequacies in data capture technology. The end goal of complex scoring systems in simulations is to obtain similar, if not better, levels of scoring as could be obtained through having experts observe candidates as they complete the simulation, but through an automatic process that allows this to occur at a fraction of the cost. This section will continue with some examples of complex scoring of automatic tasks currently being utilized as well as an overview of the process utilized to develop these complex scoring systems. Exciting opportunities for future research in this area will also be examined.

Prehire assessment simulation developers would be well-suited to delve into the training field, where increasingly realistic, complex, and autonomous simulations have made it imperative to develop new assessment methods to ensure the training was successful. One example of this can be seen in the work of Koenig et al. (2010) who developed a theoretical framework for assessing performance in a simulation based on a naval ship and then followed up their theoretical framework with a computational design that incorporated elements learned from their initial attempts (Iseli et al. 2010; Koenig et al. 2010). The authors' two reports revolving around the naval simulation and its scoring are excellent resources providing details about the development of complex scoring systems for games in which examining a person's score falls short of adequately describing their performance. It is beyond the scope of this chapter to delve into all the intricacies described by these authors; however, we will provide a brief overview of vital steps in the assessment development process for a game-based simulation.

Koenig and colleagues used a preexisting 3D simulation of a naval ship to assess situational awareness as well as knowledge of how to deal with fires and floods (Iseli et al. 2010; Koenig et al. 2010). When utilizing a preexisting scenario, the steps that must be taken to create a valid scoring approach include: (1) the use of various specification editors to determine the domain represented by the game; in essence, this step defines what is being measured at different levels in the simulation; (2) the creation of an ontology development process, involving the definition of the domain and elements within the domain as well as the creation of element equivalence classes and the definition of relations, both within and between categories of objects defined earlier; this step involves specifying the theoretical model and showing relationships between variables; (3) the generation of a Bayesian Network to create a graphical relationship representing probabilistic relationships between variables; (4) the development of analysis tools based on the Bayesian Network; and (5) the choice of

output generation tools and procedures to convey scores on pivotal aspects of the simulation to key stakeholders.

Although this type of modeling approach has most commonly been used to predict training outcomes in the past, it is hoped that as automated simulation design becomes more advanced our field may learn from these techniques. Ultimately, these techniques may help us move away from asking respondents to self-report their own psychological characteristics to enabling us to infer these characteristics directly through the decisions they make and behaviors they exhibit in the simulation. However, we do not necessarily expect modeling approaches like these to result in substantial enhancements to validity, though they may certainly add value. Rather, the primary benefit may be to better predict various latent constructs and develop our understanding of effective decision making.

5.5 Conclusion

Technological advancements can be plotted on an exponential growth curve, and in the early twenty-first century, civilization is at the knee of the curve (Kurzweil 2005). Vast amounts of information coupled with improving analytical tools will enable our predictive capabilities to asymptote. As a society, we are rapidly approaching a point where everything that can be known and understood will be. Simulations are also at the knee of their curve, and will soon peak as synergies between programming technology and the big data movement are realized.

There is substantial complexity in how simulations can be scored; and, since simulation design is evolving so rapidly, it is difficult to fully research any one technique. As our ability to simulate the real world in software evolves, we face the potential of being able to migrate from asking questions to observing behavior. This is a daunting possibility. Most of our psychometric history has been built on analyses of questions and responses and not on behavioral observations. There has simply been no automated way to observe true behavior until now. For the first time, realistic virtual scenarios are being used to simulate real life—giving job candidates the opportunity to demonstrate what they would *do* and not just what they *say* they would do.

The psychometric implications of this shift are vast. Self-report questions are conceptually simple: the researcher asks a question and the respondent answers it, resulting in a clean data point that may be linked directly to an underlying construct, which may be linked empirically and rationally to various criteria. Simulations allow us to ask job candidates to actually perform tasks—to brainstorm, multitask, respond to various stimuli, and much more. The responses they give are one type of data point; but we may also record and study their interactions with the virtual environment itself. For example, what can we learn from errant mouse clicks, repeated plays of the instructions, thorough viewing of different tabs of information in a problem-solving scenario, or response latencies on untimed exercises? In other words, we are increasingly able to capture indicators of style, and at least with our clients, we

have found many of these indicators to be excellent predictors of real-world style and results. Conceptually, this shift is important: we are no longer looking at how a theoretical construct (e.g., Extraversion) predicts an emergent outcome (e.g., sales results); instead, we are concerned with what emergent prehire behavior can teach us about a new hire's emergent posthire behavior. Moving from self-report to emergent behavior measurement constitutes a fundamental advancement, and gets researchers one step closer to real world behavior.

However, as we have discussed, a potentially larger shift in predictive power will come from the convergence of big data and combinatorial scoring. Rather than relying on one particular item or scale construct, and expecting that lone data point to predict a meaningful outcome, modern simulations allow us to easily combine diverse item types and scales, while bigger data sets allow us to more easily examine interactions among these elements. We believe that these developments will create a golden age of assessment prediction power.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free response writing tests. *Applied Psychological Measurement, 12,* 117–128.

Arthur, W. R., Day, E., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*(1), 125–154.

Attali, Y. (2004) Exploring the feedback and revision features of Criterion. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.

Bejar, I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76,* 522–532.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 67–91). Mahwah: Lawrence Erlbaum Associates.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open ended versus multiple-choice response formats—It does make a different for diagnostic purposes. *Applied Psychological Measurement, 11*(4), 385–395.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores a literature review, application to content validity, and meta-analysis. *Organizational Research Methods, 10*(4), 689–709.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park, MD.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*(1), 83–117.

Clauser, B. E., Ross, L. P., Clyman, S. G., Rose, K. M., Margolis, M. J., Nungester, R. J., Piemme, T. E., Chang, L., El-Bayoumi, G., Malakoff, G. L., & Pincetl, P. S. (1997). Development of a scoring algorithm to replace expert rating for scoring a complex performance-based assessment. *Applied Measurement in Education, 10,* 345–358.

Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer based testing: Building the foundation for future assessments* (pp. 165–192). Mahwaj: Lawrence Erlbaum Associates.

De Corte, W. (1999). Weighing job performance predictors to both maximize the quality of the selected workforce and control the level of adverse impact. *Journal of Applied Psychology, 84*(5), 695–702.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92*(5), 1380–1393.

Doverspike, D., Winter, J. L., Healy, M. C., & Barrett, G. V. (1996). Simulations as a method of illustrating the impact of differential weights on personnel selection outcomes. *Human Performance, 9*(3), 259–273.

Drasgow, F., & Olson-Buchanan, J. B. (1999). *Innovations in computerized assessment*. Mahwah: Lawrence Erlbaum Associates.

Dudley, N. M., & Cortina, J. M. (2008). Knowledge and skills that facilitate the personal support dimension of citizenship. *Journal of Applied Psychology, 93*(6), 1249–1270.

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology, 94*, 334–346.

Fielding, N. G., & Lee, R. M. (1998). *Computer analysis and qualitative research*. Thousand Oaks: Sage.

Flaugher, R. (1990). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 41–64). Hillsdale: Lawrence Erlbaum Associates.

Folk, V. G., & Smith, R. L. (2002). Models for delivery of CBTs. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer based testing: Building the foundation for future assessments* (pp. 41–66). Mahwah: Lawrence Erlbaum Associates.

Gatewood, R., Field, H. S., & Barrick, M. (2010). *Human resource selection*. Mason: Thomson South-Western.

Gibby, R. E., Biga, A., Pratt, A., &, Irwin, J. (2008). Online and unsupervised adaptive cognitive testing: lessons learned. Paper presented at the conference of the Society for Industrial and Organizational Psychology, San Francisco.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293.

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance to test systems to small-scale cheating. *International Journal of Testing, 9,* 283–309.

Hambleton, R. K. (2002). New CBT technical issues: Developing items, pretesting, test security, and item exposure. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer based testing: Building the foundation for future assessments* (pp. 193–203). Mahwah: Lawrence Erlbaum Associates.

Hanson, M. A., Borman, W. C., Mogilka, H. J., Manning, C., & Hedge, J. W. (1999). Computerized assessment of skill for a highly technical job. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 197–220). Mahwah: Lawrence Erlbaum Associates.

Hattrup, K. (2012). Using composite predictors in personnel selection. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 297–319). New York: Oxford University Press.

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self- narratives. *Journal of Research in Personality, 43*, 524–527.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer testing. *Education Measurement: Issues and Practice,* 16–25.

Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automatic assessment of complex task performance in games and simulations (CRESST Research Report No. 775)*. Los Angeles:

National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Kingsbury, G., & Weiss, D. J. (1983). A comparison of IRB-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 257–283). New York: Academic Press.

Koenig, A. D., Lee, J. J., Iseli, M., & Wainess, R. (2010). *A conceptual framework for assessing performance in games and simulations. (CRESST Research Report No. 771)*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for Studies in Education, UCLA.

Kurzweil, R. (2005). *The singularity is near*. New York: Viking.

Lee, H. S., & Cohn, L. D. (2009). Assessing Coping Strategies by Analyzing Expressive Writing Samples. *Stress and Health, 26*, 250–260.

Lord, F. M. (1970). *A theory of test scores. Psychometric monograph no. 7*. Princeton: Educational Testing Service.

Lord, F. M. (1977). A broad range tailored test of verbal ability. *Applied Psychological Measurement*, 95–100.

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

Luecht, R. M., & Clauser, B. E. (2002). Test models for complex CBT. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer based testing: Building the foundation for future assessments* (pp. 67–88). Mahwah: Lawrence Erlbaum Associates.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223–236). New York: Academic Press.

Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback–Leibler information. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer Science+Business Media.

Mumford, M. D., Zaccaro, S. J., Harding, F. D., Jacobs, T. O., & Fleishman, E. A. (2000). Leadership skills for a changing world solving complex social problems. *The Leadership Quarterly, 11*, 11–35.

Newman, D. A., Jacobs, R. R., & Bartram, D. (2007). Choosing the best method for local validity estimation: Relative accuracy of meta-analysis versus a local study versus Bayes-analysis. *Journal of Applied Psychology, 92,* 1394–1413.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351–356.

Parshall, C. G., Harmes, C., Davey, T., & Pashley, P. J. (2010). Innovative items for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer Science+Business Media.

Pyburn, K. R., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology, 61*(1), 143–151. doi:10.1111/j.1744-6570.2008.00108.x.

Rasch, G. (1960). *Probabilistic model for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment, 1*(2), 3–21.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*(3), 707–721. doi:10.1111/j.1744-6570.1997.tb00711.x.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82*(5), 719.

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

Segall, D. O. (2010). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing*. New York: Springer Science+Business Media.

Segall, D. O., & Moreno, K. E. (1999). Development of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 35–65). Mahwah: Lawrence Erlbaum Associates.

Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. In *Annual National Council on Measurement in Education Meeting* (pp. 14–16).

Smith, P. (2002). *Developing composite indicators for assessing health system efficiency. Measuring up: Improving the performance of health systems in OECD countries*. Paris: OECD.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

Tesch, R. (1990). *Qualitative research: Analysis types and software tools*. London: Farmer.

Tonidandel, S., Quinones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87,* 320–332.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22,* 195–211.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W Glas (Eds.), *Elements of adaptive testing*. New York: Springer Science+Business Media.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

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

Wainer, H., Brown, L. M., Bradlow, E. T., Wang, X., Skoupski, W. P., Boulet, J., & Mislevy, R. J. (2006). An application of testlet response theory in the scoring of a complex certification exam. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 169–199). Mahwah: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–202.

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

Zaccaro, S. J., Mumford, M. D., Connelly, M. S., Marks, M. A., & Gilbert, J. A. (2000) Assessment of leader problem-solving capabilities. *The Leadership Quarterly, 11,* 37–52.

Zickar, M. J., Overton, R. C., Taylor, L. R., & Harms, H. J. (1999). The development of a computerized selection system for computer programmers in a financial services company. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 7–33). Mahwah: Lawrence Erlbaum Associates.